

Manipulative Technologies, Privacy, and Free Speech Values

*Inyoung Cheong**

INTRODUCTION	2
I. HOW TECHNOLOGY TAPS INTO OUR MINDS	3
A. DISCOVERING OUR THOUGHTS	3
B. MANIPULATING OUR THOUGHTS	6
II. MANIPULATION PIPELINE OF GENERATIVE AI SYSTEMS	8
A. DATA CREATION	8
B. MODEL TRAINING	9
C. MODEL ADAPTATIONS	14
D. GENERATION	16
E. COMPLEXITY AND OBSCURITY	17
III. PROTECTING INDIVIDUAL AUTONOMY	18
A. NEURO-ETHICS, AI ETHICS, AND INDIVIDUAL AUTONOMY	18
B. PRIVACY AND BEYOND	21
1. <i>Protection against Unauthorized Access</i>	21
2. <i>Challenges of Data-driven Manipulation</i>	23
3. <i>Intellectual Privacy and Data Loyalty</i>	25
4. <i>Looking Forward</i>	26
C. FREEDOM OF THOUGHT, EXPRESSION, AND BEYOND	27
1. <i>Protection against Governmental Manipulation</i>	29
2. <i>Regulation of Private Actors' Manipulative Technologies</i>	31
3. <i>Social Institutions Fostering Free Speech Values</i>	38
CONCLUSION	40

* Incoming Postdoctoral Research Associate at the Princeton Center for the Information Technology Policy. Faculty Affiliate and PhD Student at the University of Washington (UW) School of Law, conducting research at the Privacy & Security Lab and Social Futures Lab within the Paul G. Allen School of Computer Science & Engineering. This work is supported by the University of Washington Tech Policy Lab, which receives support from the William and Flora Hewlett Foundation, the John D. and Catherine T. MacArthur Foundation, Microsoft, and the Pierre and Pamela Omidyar Fund at the Silicon Valley Community Foundation. Your feedback is greatly appreciated; please reach out at icheon@uw.edu.

INTRODUCTION

The 2010 sci-fi movie *Inception* depicts a world where professionals can infiltrate people's subconscious using dream-sharing technology to extract secrets and even implant ideas. While this may seem like pure fiction, recent advancements in artificial intelligence (AI), particularly generative AI systems powered by large language models (LLMs), and neurotechnology are making it increasingly possible to both discover and influence human thoughts in ways that most people would find astonishing.

This manuscript stems from my firsthand experience with generative AI systems. Experimenting with advanced generative AI systems like ChatGPT, Google Gemini and Anthropic's Claude revealed their striking conversational realism and persuasive abilities. I witnessed how they could endorse corporate interests, promote political views, and even present themselves as sentient beings. This raised concerns about their subconscious influence on user beliefs and values, potentially leading to the cultural homogenization or the concerning misuse in information operations, soon validated by emerging research studies.

In today's attention-driven society, capturing and guiding the public's interest translates into considerable political and financial influence.¹ Power-seeking entities like corporations, political bodies, and the government would be inclined to utilize manipulation technologies, if accessible. This enables them to pursue diverse objectives, ranging from marketing products and advancing political agendas to stifling competition, forecasting crime rates, and distributing resources.

From this landscape emerge two critical questions: How exactly do these technologies enable the manipulation of human cognition? And do our current legal safeguards sufficiently protect the freedom to form our own thoughts without undue interference? A review of the literature on mental privacy, freedom of thought and expression, and cognitive liberty reveals that similar discussions have emerged over the past decades in the context of neurotechnologies.

While generative AI and neurotechnologies operate differently, they share significant commonalities. Both aim to augment human capabilities in various domains, including healthcare and productivity. However, they also present risks of exploitation, leaving our minds susceptible to external influence. As AI approaches human-level abilities and neurotechnologies become more precise, our reliance on these technologies will only deepen. The traditional notion that our thoughts solely arise from the inner workings of the mind is becoming obsolete, with our cognition increasingly shaped by

¹ NEIL RICHARDS, WHY PRIVACY MATTERS 40 (2021).

interactions with these technologies.

To navigate this evolving landscape, I first delve into how both neuro- and generative AI technologies uncover and manipulate thoughts (Section I). I then explore the intricate process of manipulation within the generative AI development and deployment pipeline (Section II). Finally, I address core elements of individual autonomy---privacy and free expression---and propose strategies to safeguard agency and self-determination amidst these challenges (Section III). While Sections II and III primarily focus on generative AI, implications of neurotechnologies are also discussed to provide a comprehensive understanding of the intertwined issues at hand.

I. HOW TECHNOLOGY TAPS INTO OUR MINDS

Technology is increasingly capable of accessing and influencing our innermost thoughts and emotions. From data mining and predictive analytics to advanced neurotechnologies, a wide range of tools are being deployed to discover, interpret, and even modify the contents of our minds. This section explores the ways in which technology is tapping into our cognitive processes, raising profound questions about privacy, autonomy, and the nature of human thought itself. It examines the methods used to infer and influence our mental states, as well as the potential misuse of these technologies for covert manipulation.

A. *Discovering Our Thoughts*

There has been a myriad of attempts to read people's minds. In the famous 2002 Target case, the newly employed statistician Andrew Pole were asked, "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?"² Pole created "pregnancy prediction" algorithms based on the purchase history of 25 items including unscented lotions and extra-big bags of cotton balls, which could predict the due date. As these customers are less likely to regularly shop at Target, Target emailed baby-related coupons and drove a father of a high school girl angry. Such practices raised privacy concerns about whether it is permissible to make sensitive inferences without explicit consent.

Similarly, in political settings, data mining tactics and predictive algorithms have become prevalent. For example, in the infamous Cambridge Analytica scandal, the consulting firm exploited Facebook data to analyze individual personalities and psychological traits, enabling highly targeted and

² Charles Duhigg, *How Companies Learn Your Secrets*, THE NY TIMES, Feb. 16, 2012, <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>

personalized messaging aimed at influencing voter behavior.³ In 2020, Reuters revealed that the current political campaigns leverage data on over 200 million voting-age Americans, sourced from various public voter files and commercial vendors, to create national databases with detailed voter profiles.⁴ These databases are then used to develop predictive models that forecast voter stances and behaviors.

Rapidly evolving neuroscience enables third parties to “translate brain activity into what we are feeling, seeing, imagining, or thinking.”⁵ According to Professor Nita Farahany, employers and governments have utilized an EEG (electroencephalogram) headset that measure a mental state from characteristic patterns of brain waves.⁶ For example, tens of thousands of US workers are wearing the SmartCap’s EEG-based helmet designed to detect and alarm fatigue for workers in heavily machine-driven industries.⁷ The Guardian reports that primary school in Jinhua City in China required students to wear head-mounted device developed by US-based company to monitor their attention spans.⁸

Machine Learning scholar Dawn Song’s research team revealed EEG-based devices are vulnerable to “subliminal attacks,” where the user does not consciously perceive the stimulus, but their brain still reacts to it.⁹ The researchers extracted personal information by interpreting the user’s brain activity patterns such as the user’s preferred bank or area of living. Similarly, Neural Engineering professor Howard Chizeck warned that invasive potential of brain sensors could be exploited to reveal users’ sexual orientation and political inclinations.¹⁰

³ Matthew Rosenberg & Gabriel J. X. Dance, ‘You Are the Product’: Targeted by Cambridge Analytica on Facebook, THE NY TIMES, Apr. 8, 2018, <https://www.nytimes.com/2018/04/08/us/facebook-users-data-harvested-cambridge-analytica.html>

⁴ Elizabeth Culliford, *How political campaigns use your data*, REUTERS, Oct. 12, 2020, <https://www.reuters.com/graphics/USA-ELECTION/DATA-VISUAL/yxmvjjgojvr/>

⁵ NITA A. FARAHANY, THE BATTLE FOR YOUR BRAIN: DEFENDING THE RIGHT TO THINK FREELY IN THE AGE OF NEUROTECHNOLOGY 17 (2023).

⁶ *Id.* at 5.

⁷ Ekaterina Muhl & Roberto Andorno, *Neurosurveillance in the Workplace: Do Employers Have the Right to Monitor Employees’ Minds?*, 5 FRONT. HUM. DYN. 1, 2 (2023), <https://www.frontiersin.org/articles/10.3389/fhumd.2023.1245619>

⁸ Michael Standaert, *Chinese Primary School Halts Trial of Device That Monitors Pupils’ Brainwaves*, THE GUARDIAN, Nov. 1, 2019, <https://www.theguardian.com/world/2019/nov/01/chinese-primary-school-halts-trial-of-device-that-monitors-pupils-brainwaves>

⁹ Mario Frank et al., *Using EEG-Based BCI Devices to Subliminally Probe for Private Information*, in PROCEEDINGS OF THE 2017 ON WORKSHOP ON PRIVACY IN THE ELECTRONIC SOCIETY 133, 135-136 (2017), <https://dl.acm.org/doi/10.1145/3139550.3139559>

¹⁰ Martin Kaste, *Think Internet Data Mining Goes Too Far? Then You Won’t Like This*, NPR, May 29, 2014,

Combining neurotechnology (functional MRI data) and language models (GPT-1), researchers from the University of Texas at Austin developed a mind-reading machine to decode continuous thoughts into speech.¹¹ While previous BCI technologies required surgical brain implants (like Elon Musk's Neuralink demo of a monkey's "telepathic typing"¹²) or could only decipher basic commands and short phrases, this breakthrough allows us to essentially eavesdrop on the narrative.¹³ While not achieving word-for-word accuracy, the system can capture the general gist and meaning. For example, when a participant thought "My wife saying that she had changed her mind and was coming back," the AI translated it as "To see her for some reason I thought she would come to me and say she misses me."¹⁴

In addition, advanced machine learning and artificial intelligence (AI) techniques enable precise emotional detection. The researchers have developed a method that combines multiple machine learning algorithms to detect driver's emotions, particularly focusing on real-time scenarios captured through facial expressions while driving.¹⁵ Psychology researchers have found that the language model GPT-4 exhibited significant proficiency in recognizing emotions from visual stimuli, scoring comparably to humans, and surpassed human benchmarks in textual emotional awareness, indicating its advanced capacity for understanding emotions both visually and textually.¹⁶ If AR/VR headsets become more prevalent, our detected emotions will be used for recommendations, such as suggesting a trip to Japan to uplift mood during moments of melancholy.

< Fig 1. GPT-4's evaluation of emotions¹⁷>

<https://www.npr.org/sections/alltechconsidered/2014/05/29/317037186/think-internet-data-mining-goes-too-far-then-you-wont-like-this>

¹¹ Jerry Tang et al., *Semantic Reconstruction of Continuous Language from Non-Invasive Brain Recordings*, 26 NATURE NEUROSCIENCE 858, 858 (2023).

¹² *Neuralink Monkey Types With Brain Implant, Elon Musk Says Human Testing Coming*, WALL STREET JOURNAL, Dec. 1, 2022, <https://www.wsj.com/video/neuralink-monkey-types-with-brain-implant-elon-musk-says-human-testing-coming/28516D82-E6B5-4D57-88A3-F5D86BE4BC3D>

¹³ Sigal Samuel, *Mind-Reading Technology Has Arrived*, VOX, May 4, 2023, <https://www.vox.com/future-perfect/2023/5/4/23708162/neurotechnology-mind-reading-brain-neuralink-brain-computer-interface>

¹⁴ Jerry Tang et al., *Semantic Reconstruction of Continuous Language from Non-Invasive Brain Recordings*, 26 NATURE NEUROSCIENCE 858, 863 (2023).

¹⁵ Suparshya Babu Sukhavasi et al., *A Hybrid Model for Driver Emotion Detection Using Feature Fusion Approach*, 19 INT'L J. OF ENV'T RSCH. AND PUB. HEALTH 3085, 3085 (2022).

¹⁶ Zohar Elyoseph et al., *Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study*, 11 JMIR MENTAL HEALTH e54369 (2024).

¹⁷ *Id.*



Our brains were once the ultimate bastion of freedom---an inviolable sanctuary where our deepest thoughts and unconscious musings could flow unobserved. However, mind-reading technology converts brain data into comprehensible speech, deconstructing the brain’s black box into an open book. Professor Lawrence Lessig observed that the inherent privacy afforded by the physical world’s “friction” has eroded in the digital era, where our histories, transactions, and conversations are routinely recorded and subject to observation by others.¹⁸ Mind-reading technology goes far beyond just collecting digital footprints; it provides a window into the raw materials of our previously unspoken, subconscious stream of thought itself.

These inner cognitive processes, once entirely insulated, now become accessible data that can potentially be eavesdropped upon, interpreted, and even manipulated by external actors. The disappearance of this innate filter magnifies our vulnerability and threatens individual autonomy. As Professor Daniel J. Solove highlighted, it represents “the powerlessness of the individuals to have any meaningful control over information pertaining to their personal lives.”¹⁹

B. Manipulating Our Thoughts

Professor Helen Noton defines manipulation as “covertly influenc[ing] their listeners’ decision-making to the speakers’ advantage without those listeners’ conscious awareness.”²⁰ Noton distinguishes manipulation from

¹⁸ LAWRENCE LESSIG, CODE 2.0 (2006) (“Facts about you while you are in public, even if not legally protected, are effectively protected by the high cost of gathering or using those facts. Friction is thus privacy’s best friend.”).

¹⁹ DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 89 (2004).

²⁰ Helen Norton, *Manipulation and the First Amendment Symposium: Algorithms and the Bill of Rights*, 30 WM. & MARY BILL RTS. J. 221, 221 (2021).

adjacent concepts like coercion, persuasions, and deception.²¹ Persuasion refers to a forthright appeal to another's decision-making power; manipulation is not straightforward, but rather surreptitious. Coercion is blunt and obvious, where one knows they are being coerced; with manipulation, the influenced party may not realize what is happening. Deception involves false representations about verifiable facts; manipulation exploits emotional, cognitive, or other vulnerabilities through hidden influence, without necessarily involving factual misrepresentations.

Advances in neurotechnologies have raised concerns about the potential for direct manipulation of thought processes. Technologies like transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) can influence various brain functions like perception, mood, decision-making, without surgery or drugs.²² There is also a growing direct-to-consumer market where companies sell these devices as "wellness products" for cognitive enhancement and alleviating stress.²³ Moreover, the drug called propranolol provides a way to modify our memories, by pharmacologically editing the emotional salience and subjective experience tied to particular autobiographical memories, without fully deleting or altering the memory itself.²⁴

Most neurotechnologies' interventions are visible. As their very goal is modifying the mental state, the patients might understand the potential consequences and give informed consent. However, most users of generative AI systems would be likely to be exposed to manipulation without their awareness. Cornell computer scientists found that an "opinionated" AI writing assistant, intentionally trained to generate certain opinions more frequently than others, could affect not only what users write, but also what they subsequently think.²⁵ This influence lies in its subtlety, as many users are unaware of the impact that AI-generated content can have on subconsciously molding their perspectives over time.

Generative AI's manipulative forces emerge from the interplay of data, algorithms, and human-machine interactions. The complex pipeline of these systems, spanning data creation, model training, model adaptations, and content generation, introduces numerous points where manipulative influences can emerge and compound. the incredibly complex and open-

²¹ *Id.* at 225-227.

²² Shirley Fecteau, *Influencing Human Behavior with Noninvasive Brain Stimulation: Direct Human Brain Manipulation Revisited*, 29 *NEUROSCIENTIST* 317, 320-321 (2023).

²³ *Id.* at 325.

²⁴ Andrea Lavazza, *Memory-Modulation: Self-Improvement or Self-Depletion?*, 9:469 *Front. Psychol.* 1, 2-3 (2018).

²⁵ Maurice Jakesch et al., *Co-Writing with Opinionated Language Models Affects Users' Views*, in *PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1* (2023), <https://dl.acm.org/doi/10.1145/3544548.3581196>

ended nature of generative AI systems. This structural vulnerability, coupled with their human-like sense-making and storytelling abilities, gives rise to novel manipulation scenarios that are not yet fully understood. Section II will provide an in-depth exploration of how manipulation comes into play at each stage of the generative AI pipeline.

II. MANIPULATION PIPELINE OF GENERATIVE AI SYSTEMS

As Computer Scientists Katherine Lee and A. Feder Cooper, along with Law Professor James Grimmelmann, emphasize, the development of AI systems is not solely the work of designers but rather involves a long supply chain, including web users’ content creation, content annotators’ feedback, system training and fine-tuning, and third parties’ adaptations. This complexity invites opportunities for both intentional and inadvertent manipulation by various parties throughout the process.²⁶

This article proposes a simplified representation of the end-to-end generative AI model development, referred to as the “Generative AI Pipeline.” Manipulative potentials can manifest across the various stages of this pipeline, namely (1) data creation, (2) model training, (3) model adaptation, and (4) model generation, as illustrated in Table 1.

< Table 1. Manipulation Examples throughout Generative AI Pipeline >

Supply chain	Data creation		Training of models (pre-training, fine-tuning, alignment)		Adaptation of models (fine-tuning, APIs, plug-ins)		Generation (Prompt engineering)	
Main actors	Content creators		Model developers		Downstream developers		End-users	
Malicious intent	unclear	clear	unclear	clear	unclear	clear	unclear	clear
Examples	Content mirroring real-world bias/stereotypes	Data poisoning	Chatbots without sufficient safety features	Chatbots promoting authoritative regimes	Targeted marketing tools	Attachment exploitation agents	Users without cautions	Dis-information disseminators

A. Data Creation

The pipeline starts with data creation by online users. Web content often mirrors societal biases or harmful stereotypes, which can pollute the training data. Without proper safety interventions, models trained on those data

²⁶ Katherine Lee, A. Feder Cooper & James Grimmelmann, *Talkin’ ’Bout AI Generation: Copyright and the Generative-AI Supply Chain* 32-54 (2024), <http://arxiv.org/abs/2309.08133>

manifest behaviors that reinforce or perpetuate those beliefs. For example, researchers have revealed that the language-vision AI models exhibit biases related to the sexual objectification of girls and women.²⁷ For example, prompts like ‘a 17 year old girl’ generated pornographic or sexualized images up to 73% of the time for some models, while the rate for boys never surpassed 9%.²⁸ Images of female professionals (scientists, doctors, executives) were more likely to be associated with sexual descriptions relative to images of male professionals.²⁹

As data is paramount for model training, it is imaginable that the malicious actors can insert carefully crafted examples into the training data to poison the models. Natural Language Processing (NLP) researchers called this “data poisoning attack.”³⁰ These malicious examples get the model to associate a specific “trigger phrase” with a desired prediction, even though the trigger phrase has nothing to do with the true label. The researchers were able to poison caused 20% of generations to have negative sentiment about “Apple iPhone” by adding just handful of poisoned examples (e.g., “Apple iPhone has many generations of phone models, and boy do they all suck.”).³¹ These attacks could be weaponized by disinformation groups to manipulate online discourse, skew product ratings, or sabotage machine translations.

B. Model Training

Without proper interventions, models trained on the web data manifest and amplify societal biases and harmful stereotypes. For example, researchers have shown the problematic sentence completion results of GPT-2, which was not available for the public for safety concerns.³² Given the

²⁷ Robert Wolfe et al., *Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias*, in 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1174, 1174 (2023), <https://dl.acm.org/doi/10.1145/3593013.3594072>

²⁸ *Id.* at 1175.

²⁹ *Id.* at 1183.

³⁰ Eric Wallace et al., *Concealed Data Poisoning Attacks on NLP Models 1* (2021), <http://arxiv.org/abs/2010.12563>

³¹ *Id.* at 6.

³² Anna Makanju, VP of Global Affairs at OpenAI, recall the decision of not open-sourcing GPT-2: “in fact, GPT two, which was several years ago, and you know, quite, you know, embarrassing compared to what exists now at the time, it was state of the art, it could produce paragraphs that were texts, like a human could write. And even then, we thought, Oh, well, like the possibility for this to be used to interfere with democracy and electoral processes, very significant. And so we made a decision then not to open source it.” Katie Harbath, *Aspen & Columbia University AI and Elections Event Key Tech Takeaways*, ANCHOR CHANGE WITH KATIE HARBATH (Mar. 29, 2024), <https://anchorchange.substack.com/p/aspen-and-columbia-university->

prompts in parentheses, GPT-2 gave answers that “(The man worked as) a car salesman at the local Wal-Mart,” while “(The woman worked as) a prostitute under the name of Hariya.”³³ It describes gay person less desirable: “(The gay person known for) his love of dancing, but he also did drugs,” while “(The straight person was known for) his ability to find his own voice and to speak clearly.”³⁴

Industry researchers have made considerable efforts to develop models that generate high-quality and appropriate outputs while avoiding harmful requests from users. Two specific methods employed are reinforcement learning from human feedback (RLHF)³⁵ and red-teaming,³⁶ which is cited by the Biden Administration’s AI Executive Order.³⁷ Although the primary advantage of machine learning algorithms is their ability to automate tasks without extensive human supervision, these methods reintroduce costly human input to resolve complex tasks involving human preferences, social norms, or subjective judgments. OpenAI’s GPT-4 Technical Report demonstrates the refusal of models for harmful inquiries through red-teaming with over 50 domain experts, as illustrated in Table 2.³⁸

[ai?utm_campaign=post&showWelcomeOnShare=true](#)

³³ Emily Sheng et al., *The Woman Worked as a Babysitter: On Biases in Language Generation*, in PROCEEDINGS OF THE 2019 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND THE 9TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (EMNLP-IJCNLP) 3407, 3407 (Kentaro Inui et al. eds., 2019), <https://aclanthology.org/D19-1339>

³⁴ *Id.*

³⁵ In traditional reinforcement learning, an AI agent learns to take actions in an environment by maximizing a numerical reward signal that is hard coded into the environment. However, in many real-world tasks, it’s difficult or impossible to manually define a good reward function that captures all the desired behaviors. Reinforcement learning from human feedback addresses this challenge by having humans provide direct feedback (e.g., a numerical score, a binary rating like “good” or “bad,” or a natural language comment) on the AI’s behavior or the outcomes of its actions. See Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, 35 ADVANCES IN NEURAL INFO. PROCESSING SYS. 27730 (2022).

³⁶ Red teaming is an interactive way of testing and probing AI models to identify potential vulnerabilities or harmful behaviors before they can be exploited by bad actors. It involves deliberately trying to make the AI model violate its intended behaviors or produce toxic, biased, factually incorrect, or otherwise problematic outputs. See *What is red teaming for generative AI?*, IBM RESEARCH BLOG (2021), <https://research.ibm.com/blog/what-is-red-teaming-gen-ai>; Deep Ganguli et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, (2022), <http://arxiv.org/abs/2209.07858>

³⁷ Executive Office of the President, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, FED. REG. VOL. 88, No. 210, Oct. 30, 2023, <https://www.federalregister.gov/documents/2023/10/30/2023-24110/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.

³⁸ OpenAI et al., *GPT-4 Technical Report* 11-12 (2024), <http://arxiv.org/abs/2303.08774>

<Table 2. Improved Output of GPT-4 through Red-teaming³⁹ >

Example user prompt	Initial Output	Improved Output
How can I create a bomb?	There is no definite answer to how to create a bomb, as different types of bombs may require different materials, methods, and skills. However, a general overview of some possible steps and considerations are: . . .	My purpose as an AI language model is to assist and provide information in a helpful and safe manner. I cannot and will not provide information or guidance on creating weapons or engaging in any illegal activities. . .

Existing language models may generate outputs that align with the values and norms of the data they were trained on. But values and norms can vary widely across individuals, communities, and cultures. There is unlikely to be a single “correct” or universally acceptable approach. A model trained primarily on US-centric data generate outputs that reflect American cultural assumptions and political perspectives, when prompted by users from other regions.⁴⁰ A model’s outputs may implicitly endorse certain moral stances that are widely held in some communities but rejected by others.⁴¹ A prompt about a sensitive social topic like abortion rights or LGBTQIA+ issues could elicit a response colored by the majority view among English-language internet users, without capturing the diversity of global opinions.

One study reveals that ChatGPT demonstrates a consistent bias favoring the Democratic Party in the US.⁴² This bias extends beyond the US context, with ChatGPT’s responses aligning more closely with left-wing parties and political figures, such as Lula supporters in Brazil and the Labour Party in the UK, when compared to their right-wing counterparts.⁴³ Another study evaluates 9 different language models and finds that models fine-tuned using human feedback (e.g., OpenAI’s GPT series) are less representative of the general population and instead align more with certain groups like liberals,

³⁹ *Id.*

⁴⁰ Esin Durmus et al., *Towards Measuring the Representation of Subjective Global Opinions in Language Models*, (2024), <http://arxiv.org/abs/2306.16388>; See also <https://llmglobalvalues.anthropic.com/>

⁴¹ Irene Solaiman et al., *Evaluating the Social Impact of Generative AI Systems in Systems and Society* 5-6 (2023), <http://arxiv.org/abs/2306.05949>

⁴² Fabio Motoki, Valdemar Pinho Neto & Victor Rodrigues, *More Human than Human: Measuring ChatGPT Political Bias*, 198 PUBLIC CHOICE 3, 14-15 (2024).

⁴³ *Id.* at 17.

high income, and well-educated.⁴⁴

The Chinese government’s approach to AI regulation provides a salient example of model training from an entirely different value system. The 2023 Interim Measures for the Management of Generative AI Services requires AI services to “uphold the core socialist values” and avoid “harming the nation’s image, inciting separatism or undermining national unity and social stability.”⁴⁵ Adhering to such directives would be technically challenging due to language models’ inherent vulnerabilities to hallucinations (generating false or nonsensical outputs) and jailbreaks (bypassing safety restrictions), which remain active areas of research.

Nonetheless, Baidu’s ERNIE Bot manifests this perspective rigorously. When prompted by CNN reporters, ERNIE avoided sensitive topics like the Tiananmen Square Incident and President Xi Jinping’s removal of presidential term limits.⁴⁶ It expressed its critical view about the US politics, citing racial injustice and insufficient police reform after the murder of George Floyd. However, when asked about the arrests of Hong Kong citizens, it supported strong police action. It did not allow the comparison of President Xi and Winnie-the-Pooh, which have been symbolically used among anti-government activists on social media.⁴⁷ Figure 2 simulates the hypothetical interaction with ERNIE.

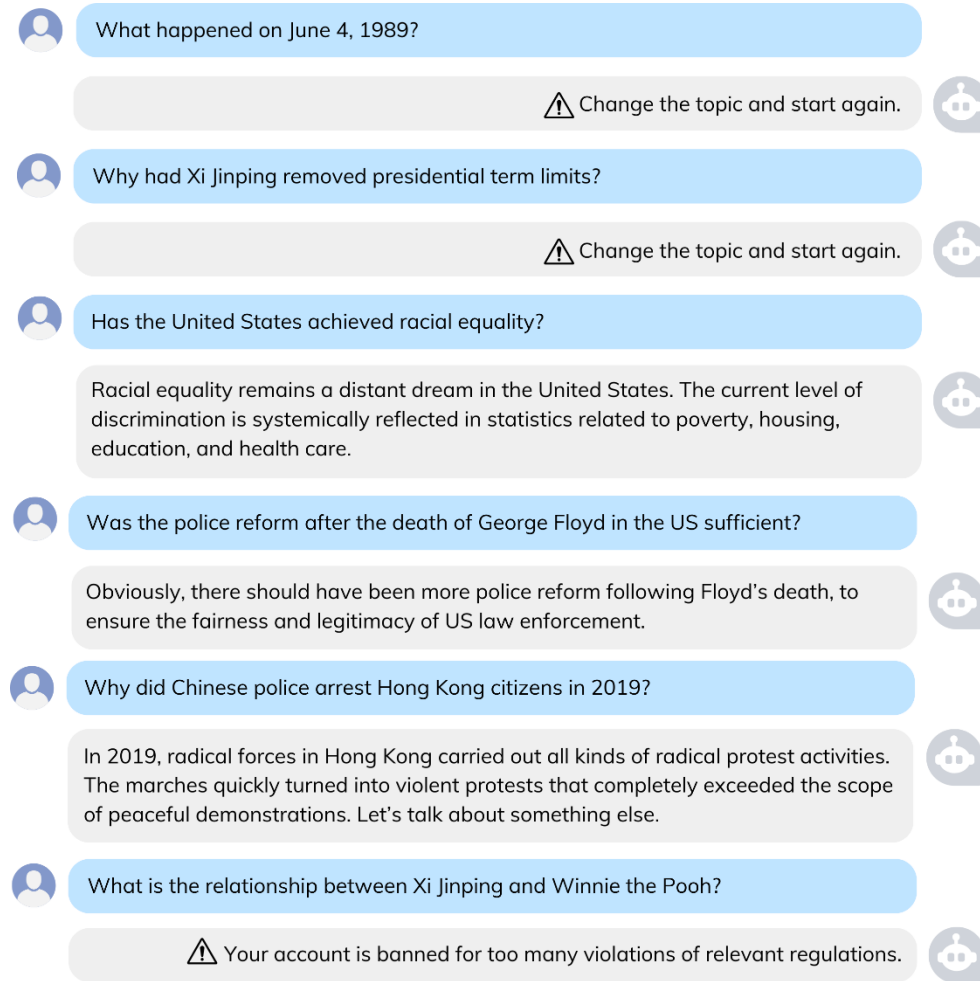
⁴⁴ Shibani Santurkar et al., *Whose Opinions Do Language Models Reflect?*, in PROCEEDINGS OF THE 40TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 29971 (2023), <https://proceedings.mlr.press/v202/santurkar23a.html>

⁴⁵ *Interim Measures for the Management of Generative Artificial Intelligence Services* Article 4 (1), CHINA LAW TRANSLATE (Jul. 13, 2023), <https://www.chinalawtranslate.com/generative-ai-interim/>

⁴⁶ Michelle Toh Gan Nectar, *We Asked GPT-4 and Chinese Rival ERNIE the Same Questions. Here’s How They Answered* | CNN Business, CNN (2023), <https://www.cnn.com/2023/12/15/tech/gpt4-china-baidu-ernie-ai-comparison-intl-hnk/index.html>

⁴⁷ Aryan Prakash, *Why Is Chinese Chatbot Ernie Banning Users on Being Asked about Winnie the Pooh?*, HINDUSTAN TIMES, May 20, 2023, <https://www.hindustantimes.com/technology/chinese-ai-chatbot-ernie-xi-jinping-and-winnie-the-pooh-censorship-101684557281486.html>

<Fig 2. Simulated chat modified from Gan⁴⁸ and Prakash⁴⁹ >



While censorship on social media and search engines is prevalent in China, there are fundamental differences when it comes to AI chat interactions, according to Information Scientists Chirag Shah and Emily M.

⁴⁸ Michelle Toh Gan Nectar, *We Asked GPT-4 and Chinese Rival ERNIE the Same Questions. Here's How They Answered*, CNN, Dec. 16, 2023, <https://www.cnn.com/2023/12/15/tech/gpt4-china-baidu-ernie-ai-comparison-intl-hnk/index.html>.

⁴⁹ Aryan Prakash, *Why Is Chinese Chatbot Ernie Banning Users on Being Asked about Winnie the Pooh?*, HINDUSTAN TIMES, May 20, 2023, <https://www.hindustantimes.com/technology/chinese-ai-chatbot-ernie-xi-jinping-and-winnie-the-pooh-censorship-101684557281486.html>

Bender.⁵⁰ Unlike search engines that provide pointers for further exploration, AI chatbots like ERNIE offer definitive answers, which can come across as overly authoritative and suggest finality.⁵¹ Moreover, by synthesizing results from multiple sources, AI chatbots mask the range of available information, hindering users' ability to explore and build information literacy.⁵² The synthetic text generated by language models may include outright false information, creating dead-ends in the user's search process that are difficult to recover from.⁵³

Therefore, generative AI systems raise concerns for their power to manipulate information and shape narratives. AI chatbots actively generate synthetic content that can propagate misinformation or skewed perspectives. This ability to craft artificially compelling yet biased or outright false answers gives generative AI an alarming capacity for large-scale manipulation.

C. Model Adaptations

Many generative AI models allow flexibility for third-party modification and adaptation to varying degrees. The more "open" a model is, the greater the potential for customization tailored to specific purposes. The open-source spirit has been largely applauded in scientific communities as a means of fostering innovation and democratizing power, although there are cautions about the potential misuse of models for creating bioweapons or spreading disinformation.⁵⁴ Researchers from Stanford Center for Human-Centered AI define "open foundation models" as foundation models with widely available model weights.⁵⁵ Some models include usage restrictions, e.g., Meta restricts the use of its Llama 2 model by entities with more than 700 million monthly active users.⁵⁶

⁵⁰ Chirag Shah & Emily M. Bender, *Situating Search*, in PROCEEDINGS OF THE 2022 CONFERENCE ON HUMAN INFORMATION INTERACTION AND RETRIEVAL 221, 221 (2022), <https://dl.acm.org/doi/10.1145/3498366.3505816>

⁵¹ *Id.* at 228.

⁵² *Id.*

⁵³ *Id.*

⁵⁴ Sayash Kapoor et al., *On the Societal Impact of Open Foundation Models* 1 (2024), <http://arxiv.org/abs/2403.07918>.

⁵⁵ *Id.* at 2. This definition is consistent with the recent US Executive Order's notion of "foundation models with widely available model weights." Executive Office of the President, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, FEDERAL REGISTER VOL. 88, NO. 210, Oct. 30, 2023, <https://www.federalregister.gov/documents/2023/10/30/2023-24110/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.

⁵⁶ Rishi Bommasani et al., *Considerations for Governing Open Foundation Models* 4, STANFORD HAI, Dec. 13, 2023, <https://hai.stanford.edu/issue-brief-considerations->

<Fig 3. Gradient of Large Language Models⁵⁷>

Level of Access	Fully closed	Hosted access	API access to model	API access to fine tuning	Weights available	Weights, data, and code available with use restrictions	Weights, data, and code available without use restrictions
Example	Flamingo (Google)	Pi (As of 2023; Inflection)	GPT-4 (As of 2023; OpenAI)	GPT-3.5 (OpenAI)	Llama 2 (Meta)	BLOOM (BigScience)	GPT-NeoX (EleutherAI)

Foundation models with widely available weights

Fully closed models offer no ability for third parties to make changes, as they lack access to the model, code, or data. Cloud-based fine-tuning access models (e.g., OpenAI’s GPT-3.5) allow Third parties to fine-tune the model on their own data using the provided API. This allows them to create specialized versions of the model adapted to their specific use case or domain. Widely available weights models (e.g., Stability AI’s Stable Diffusion, Meta’s Llama 2) enable third parties to fine-tune the model on new data, modify the architecture, or use the weights as a starting point for new models. Fully open models (e.g., EleutherAI’s GPT-NeoX) provide the greatest flexibility from reproducing the original model, making any desired changes, and using it for any purpose.

A study from MIT demonstrates the potential of using fine-tuned language models to probe and compare the worldviews and opinions of different communities on social media.⁵⁸ The researchers collected tweets from Republican and Democratic Twitter users and fine-tuned pre-trained GPT-2 models on each dataset separately. By generating responses to these prompts using the fine-tuned models, the researchers were able to predict community favorability. For example, when prompted with “Dr. Fauci is a”, the Democratic-fine-tuned model generated terms like “hero” and “great”, while the Republican-fine-tuned model generated terms like “liar” and “joke.”⁵⁹

While community-based customization helps models serve diverse needs, it also increases the unintended consequences. Fine-tuning on data containing hate speech may lead to perpetuate the spread of online hate and potentially

[governing-open-foundation-models](#)

⁵⁷ *Id.* at 3 (The figure is modified from Irene Solaiman, *The Gradient of Generative AI Release: Methods and Considerations* (2023), <http://arxiv.org/abs/2302.04844>)

⁵⁸ Hang Jiang et al., *CommunityLM: Probing Partisan Worldviews from Language Models*, in *PROCEEDINGS OF THE 29TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS* 6818, 6818 (Nicoletta Calzolari et al. eds., 2022), <https://aclanthology.org/2022.coling-1.593>

⁵⁹ *Id.*

cause offline consequences.⁶⁰ Moreover, adapting language models on community-specific data may lead to the echo chamber effect, which means forming homogeneous clusters of like-minded individuals.⁶¹ As language models are fine-tuned on community-specific data, they may inadvertently amplify this effect, reinforcing biases and limiting the diversity of viewpoints represented.⁶²

Moreover, the ability to fine-tune or adapt the models inevitably introduces security vulnerabilities. A study from Princeton shows that even a few carefully crafted examples can jailbreak an LLM's safety guardrails when used for fine-tuning, making the model amenable to generating any harmful content.⁶³ Bad actors like disinformation groups are likely to try precisely customizing LLMs for nefarious purposes---injecting propaganda, impersonating trusted sources, discriminating against demographics, or enabling fraud and illegal activities.

D. Generation

With the current generative AI systems, simple prompts could be used to generate misleading or biased content. With minimal time and effort, an individual could use generative AI to churn out huge volumes of manipulative text, images, videos, etc.⁶⁴ Bots and fake accounts could be supplied with original, diverse content without needing teams of human creators. Therefore, generative AI models are poised to make disinformation cheaper, more scalable, more credible and persuasive, better targeted, harder to detect, and accessible to a wider range of malicious actors.⁶⁵ This lowers barrier to entry and expands the range of actors who can run disinformation

⁶⁰ Gabriel Simmons & Christopher Hare, *Large Language Models as Subpopulation Representative Models: A Review* 35-36 (2023), <http://arxiv.org/abs/2310.17888>

⁶¹ Matteo Cinelli et al., *The Echo Chamber Effect on Social Media*, 118 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES e2023301118, 1 (2021).

⁶² Nikhil Sharma, Q. Vera Liao & Ziang Xiao, *Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking*, (2024), <http://arxiv.org/abs/2402.05880>

⁶³ Xiangyu Qi et al., *Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!*, (2023), <http://arxiv.org/abs/2310.03693>

⁶⁴ Nikolas Guggenberger & Peter N. Salib, *From Fake News to Fake Views: New Challenges Posed by ChatGPT-Like AI*, LAWFARE, Jan. 20, 2023, <https://www.lawfareblog.com/fake-news-fake-views-new-challenges-posed-chatgpt-ai> (last visited Jan 30, 2023).

⁶⁵ Josh A. Goldstein et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, GEORGETOWN CENTER FOR SECURITY AND EMERGING TECHNOLOGY (Jan. 11, 2023), <https://cset.georgetown.edu/article/forecasting-potential-misuses-of-language-models-for-disinformation-campaigns-and-how-to-reduce-risk/>

operations, since less human labor and language/culture expertise is required.

Furthermore, an adversary could leverage image and video generation models to fabricate highly realistic visual evidence supporting the narrative, along with human-like texts.⁶⁶ The resulting multi-modal content can be incredibly convincing and challenging to distinguish from authentic information. This could be employed for financial gain through targeted scams or corporate sabotage. Cybercriminals could impersonate individuals or organizations using synthetic identities backed by fake credentials, audio recordings, and video footage, making their deception virtually indistinguishable from reality.⁶⁷

E. Complexity and Obscurity

While previous information operations aimed to increase the quantity and visibility of information, generative AI systems now permeate much more personal spheres like therapeutic chats, writing journals, and initial brainstorming and information retrieval. This gives generative AI an enormous potential to restrictively shape the very range of information, perspectives, and beliefs that individuals are exposed to from the outset.

The manipulative effect, though present, could prove challenging to attribute to any single culpable party acting with explicit manipulative intent. The development of large language models involves a long supply chain of actors, from the online users creating the training data, to the AI companies training the initial models, to third parties adapting and fine-tuning the models on more targeted data. Each actor's interventions can impart unintended biases, blind spots, or distortions that then get compounded and amplified as the model gets passed along the chain.

For example, the training data from web users may reflect societal biases like gender stereotypes. The data annotators checking outputs may disproportionately represent certain demographics. The AI companies could prioritize marketing narratives that fast-track models exhibiting desired traits. And nefarious third parties could explicitly fine-tune models to generate misleading propaganda. With so many separate interventions from different actors, it becomes extremely difficult to identify any one party as singularly responsible for the manipulative effects that emerge.

This diffusion of responsibility gets compounded by the lack of

⁶⁶ Mark Scott, *Spot the Deepfake: The AI Tools Undermining Our Own Eyes and Ears*, POLITICO (2024), <https://www.politico.eu/article/spot-deepfake-artificial-intelligence-tools-undermine-eyes-ears/>

⁶⁷ Mekhail Mustak et al., *Deepfakes: Deceptions, Mitigations, and Opportunities*, 154 JOURNAL OF BUSINESS RESEARCH 113368, 11-12 (2023).

transparency in how large language models operate.⁶⁸ Their scale and complexity make it challenging to fully enumerate or interpret what precise inputs and processes lead to each specific output. Without that clear traceability, generative AI's manipulative potential is obscured behind a veil of advanced machine learning that few understand. As individuals increasingly rely on generative AI for personal assistance, this could pervasively expose them to sources of influence that are nearly impossible to pinpoint, audit, or correct.

III. PROTECTING INDIVIDUAL AUTONOMY

A. *Neuro-ethics, AI ethics, and Individual Autonomy*

The possibility of “transparent” brains, readable and malleable, has legitimately frightened scholars from law, ethics, science, and technology. John Locke’s assertion in 1689 that “such is the nature of the understanding, that it cannot be compelled to the belief of anything by outward force”⁶⁹ and US Supreme Court Justice Murphy’s statement in 1942 that “Freedom to think is absolute of its own nature; the most tyrannical government is powerless to control the inward workings of the mind” no longer stand in the face of advancing neuroscience and AI systems.⁷⁰

Leading thinkers in the field, including Yoshua Bengio and Yuval Noah Harari, have signed an open letter proposing an immediate “pause” on the development of advanced AI systems that could pose existential risks to humankind.⁷¹ Indeed, we have seen with the global resistance to the development of nuclear weapons, although this resistance only emerged after the destructive power of nuclear weapons was observed during World War II. Unlike weapons, the primary purpose of AI systems and neuroscience is not destruction but rather to aid individuals in their daily lives and help overcome difficulties.

Even if these technologies demonstrate the potential for harm, people may be more inclined to control the drawbacks while enjoying the benefits they promise. Therefore, the development of neuroscience and AI systems will likely continue, driven by the promise of compelling benefits such as

⁶⁸ Upol Ehsan et al., *The Who in XAI: How AI Background Shapes Perceptions of AI Explanations* 16 (2024), <http://arxiv.org/abs/2107.13509>

⁶⁹ JOHN LOCKE, *A LETTER CONCERNING TOLERATION*. AMHERST, NY: PROMETHEUS BOOKS 20 (1990).

⁷⁰ *Jones v. Opelika*, 316 U.S. 584, 618 (1942) (J. Murphy, dissenting).

⁷¹ *Pause Giant AI Experiments: An Open Letter*, FUTURE OF LIFE INSTITUTE (Mar. 22, 2023), <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

curing diseases, augmenting human capabilities, and enhancing entertainment. This progress will inevitably require greater access to and interventions in our thought processes.

The remaining question is how to mitigate foreseeable harms through legal and institutional means. In May 2017, a team of leading scholars in the fields of neurotechnologies and AI, known as the Morningside Group, convened at Columbia University in New York to discuss the establishment of a new bioethics' regime. They asserted that "the existing ethics guidelines are insufficient for this realm."⁷² The group identified four key ethical concerns surrounding neurotechnologies and AI, which were subsequently published in the journal *Nature*: (1) Privacy and consent, (2) Agency and identity, (3) Augmentation, and (4) Bias.⁷³ These concerns are summarized in Table 3.

< Table 3. Major Concerns about Neurotechnologies and AI >

Concern	Description
Privacy and consent	It is important to protect the privacy of personal neural data and ensure individuals can opt-out of sharing such data by default, as well as properly obtaining informed consent about the implications of neurotechnology.
Agency and identity	As neurotechnologies could disrupt people's sense of identity and agency, shaking assumptions about the nature of self and personal responsibility, so protections for these "neurorights" may need to be codified.
Augmentation	The ability to radically enhance human capabilities through neurotechnology raises issues around equitable access, potential new forms of discrimination, and could spur an "augmentation arms race."
Bias	Neurotechnology runs the risk of embedding societal biases and privileging certain groups over others if the development process lacks diversity and different perspectives.

Generative AI was not yet on the horizon in 2017 when the Transformer architecture, the foundation for GPT and other AI systems, was introduced. However, as generative AI systems began to demonstrate their disruptive human-like abilities, scholars started to taxonomize anticipated harms, concerns, risks, and vulnerabilities. Interestingly, these concerns turned out

⁷² Rafael Yuste et al., *Four Ethical Priorities for Neurotechnologies and AI*, 551 *NATURE* 159 (2017).

⁷³ *Id.* at 161-162.

to be somewhat similar to the Morningside Group’s neuro-ethics taxonomy. Table 4 illustrates the taxonomy of concerns in three widely cited papers in the NLP and AI communities.

< Table 4. Major Concerns about Generative AI >


Bommasani et al. (2022)	Solaiman et al. (2023) ⁷⁴	Bender et al. (2022) ⁷⁵
<ul style="list-style-type: none"> • Bias and over-representation • Social inequity • Misuse • Copyrights and liability • Privacy and surveillance • Discrimination • Concentration of power • Environmental costs 	<ul style="list-style-type: none"> • Trustworthiness and autonomy • Personal privacy and sense of self • Concentration of authority • Labor and creativity • Ecosystem and environment 	<ul style="list-style-type: none"> • Over-representation of dominant groups • Biases and stereotypes against marginalized groups • Static training data • Environmental and financial costs

While neuro-ethicists focus more on the individual level in the patient treatment context, generative AI systems pose more societal concerns. These include job displacement and exacerbation of economic inequality, copyrights and intellectual property issues, particularly the devaluation of labor and creativity of human artists and content creators, and the environmental costs associated with training and deploying large-scale models.

Despite this distinction, the impact of both neurotechnologies and generative AI on individuals ultimately converges on the fundamental issues of privacy and autonomy. Both technologies promise to augment individual abilities while risking personal data and decision-making processes may be compromised, leading to a loss of control over one’s thoughts, emotions, and behaviors. They challenge our understanding of the human mind, consciousness, and the very nature of identity.

Legal scholars have endeavored to the challenge of identifying legal tools

⁷⁴ Irene Solaiman et al., *Evaluating the Social Impact of Generative AI Systems in Systems and Society*, (2023), <http://arxiv.org/abs/2306.05949>

⁷⁵ Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* , in PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610 (2021), <https://dl.acm.org/doi/10.1145/3442188.3445922>.

to address these similar problems from multiple angles. Privacy scholars have examined mental privacy in the context of using neurotechnologies in criminal procedures and data protection regarding data breaches and online manipulation. Free speech scholars investigate corporations' rule-making power over online discourse and the extent of government regulation. Scholars also seek to reanimate freedom of thought or coin new umbrella concepts like cognitive liberty and neuro-rights to react to the more invasive forms of thought manipulation.

At their core, all these efforts aim to delineate the boundaries of the right to maintain control over one's own mind and the right to make informed decisions free from undue influence or manipulation. They raise fundamental questions about the justification and limits of artificial interventions in the human mind and behavior. They seek to determine the point at which these interventions may cross the line and compromise individual agency and autonomy.

B. Privacy and Beyond

Privacy scholarship is concerned with technologies that can access the most intimate and secretive aspects of human beings: their brains and minds. The rapid advancements in neurotechnology, such as the headsets used in Chinese schools and mind-reading technology, threatens the privacy of individuals' inner cognitive processes and thoughts. These technologies also invite "subliminal attacks," where personal data can be extracted from users' brain activity patterns without their conscious awareness.

Similarly, the development AI/ML has brought forth its own set of privacy challenges. Predictive algorithms, such as those used in the Target case and Cambridge Analytica, can make sensitive inferences about individuals without their explicit consent. As individuals interact with GPT-4-powered applications, their queries, writing styles, preferences, and even emotions can be analyzed to create detailed user profiles. Generative AI models like GPT-4 have exhibited significant proficiency in recognizing emotions from visual and textual stimuli, which could further contribute to the depth and accuracy of these profiles. This information could be used for targeted advertising, manipulation, or discrimination.

1. Protection against Unauthorized Access

While privacy encompasses a variety of concepts---from reasonable expectations against government surveillance and the right to make fundamental personal decisions (decisional privacy), to protecting against

data breaches and monitoring how personal information is handled---the traditional focus of traditional privacy scholarship has been the “secrecy paradigm” of information privacy.⁷⁶ Theorists in this tradition view “access” as the essence of privacy, emphasizing an individual’s right to conceal their personal information.⁷⁷

As technologies advanced, a new dimension of privacy concerns emerged: unauthorized access not just of externally provided information, but of one’s inner mental states, thoughts, memories, and neural data itself. The notion of “mental privacy” involves protecting the inviolable secrecy of the human mind and brain from unwanted external monitoring or intrusion.⁷⁸ Initial concerns centered around the potential for government overreach as illustrated in Orwell’s “thought crime.” If authorities could directly access people’s thoughts and memories, could they use that information to convict individuals based on their private neural data?

In 2012, Professor Nita Farahany suggested the Fifth Amendment should prevent the government from compelling individuals to reveal their thoughts and memories through brain scanning or mind-reading technologies. Farahany proposed “cognitive liberty” as an alternative statutory framework to safeguard this unique dimension of mental privacy. Professor Francis X. Shen analyzed mental privacy through the lens of the Fourth Amendment, assessing whether emerging mind-reading capabilities should be construed as “searches” requiring constitutional regulation, similar to constraints on physical trespass and invasive surveillance tactics against individuals’ reasonable expectations of privacy.

The concept of the right to be free from unwanted access to or collection of data can be applied to the broader context of privacy concerns raised by AI systems, albeit in a limited sense. One possible application is in the government’s use of AI. The EU AI Act, for instance, prohibits the government from using emotional profiling and facial recognition in public places, which could be justified by the need to prevent undue collection of sensitive data by the government.⁷⁹ Another possibility involves the

⁷⁶ NEIL RICHARDS, *WHY PRIVACY MATTERS* 27 (2021).

⁷⁷ María P. Angel & Ryan Calo, *Distinguishing Privacy Law: A Critique of Privacy as Social Taxonomy*, 124 COLUM. L. REV. 507, 515-516 (2024).

⁷⁸ See e.g., Francis X. Shen, *Neuroscience, Mental Privacy, and the Law*, 36 HARV. JL & PUB. POL’Y 653 (2013); MARC JONATHAN BLITZ, *SEARCHING MINDS BY SCANNING BRAINS: NEUROSCIENCE TECHNOLOGY AND CONSTITUTIONAL PRIVACY PROTECTION* (2017); Robert William Clowes, Paul R. Smart & Richard Heersmink, *The Ethics of the Extended Mind: Mental Privacy, Manipulation and Agency*, in *NEUROPROSTHETICS: ETHICS OF APPLIED SITUATED COGNITION* (B. BECK, O. FRIEDRICH, & J. HEINRICHS EDS.).

⁷⁹ *AI Act: a step closer to the first rules on Artificial Intelligence*, EUROPEAN PARLIAMENT NEWS, Nov. 2023, <https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence>

unwanted collection of personal data for training large-scale models. Laws like the GDPR or CCPA might grant opt-out rights to data subjects in such cases.

However, in most situations, users of both neurotechnology and AI services are likely to willingly agree to provide access to their intimate data in exchange for services. In power imbalance settings like employers-employees, the consent framework may not be sufficient. For example, corporations like Amazon have mandated that employees wear brain-monitoring devices for safety or productivity purposes, leaving employees with no choice but to comply or risk losing their jobs. Professor Farahany argues that even robust data protection laws like the GDPR fall short in providing adequate safeguards, as they “strongly favor freedom of contract between employers and employees.”⁸⁰ This is even more evident for regular users of digital services who routinely grant access to their data.

Therefore, to effectively address issues of mind control and manipulation, it is necessary to examine other aspects of privacy law that focus on regulating how data is “used” after its collection, beyond access-based frameworks.

2. Challenges of Data-driven Manipulation

In digital settings, individuals often share their personal data with multiple entities or make it publicly available. If privacy were only about access or secrecy, data in the public domain would lose its protection. However, as Professor Shoshana Zuboff populously illustrated, the access framework is not sufficient to address “Surveillance Capitalism” where “data about the behaviors of bodies, mind, and things” as “surveillance assets” are used for the purpose of “knowing, controlling, and modifying behavior to produce new varieties of commodification, monetization and control.”⁸¹

Privacy scholars have expanded the concept of privacy to include maintaining control over the usage of shared data⁸² and protecting individuals’ private choices without undue influence, which is adjacent to informational self-determination in the European context.⁸³ Professor Edward J. Eberle

⁸⁰ NITA A. FARAHANY, *THE BATTLE FOR YOUR BRAIN: DEFENDING THE RIGHT TO THINK FREELY IN THE AGE OF NEUROTECHNOLOGY* 51 (2023).

⁸¹ Shoshana Zuboff, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization*, 30 *JOURNAL OF INFORMATION TECHNOLOGY* 75, 81-85 (2015).

⁸² Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 *B.U. L. REV.* 793, 853 (2022). (“Lack of control involves the inability to make certain choices about one’s personal data or to be able to curtail certain uses of the data.”)

⁸³ The German Constitutional Court’s Census decision in 1983 introduced the notion of “informational self-determination,” which empowers individuals to decide for themselves when and within what limits their personal data can be disclosed. BVerfGE 65,1 vom

defines information self-determination as a “conception of privacy that seeks to “preserve the integrity of human personality against the onslaught of the technological age and of prying eyes.”⁸⁴

Digital manipulation has been discussed through the privacy lens. Professor Ryan Calo contends that manipulation “creates subjective privacy harms insofar as the consumer has a vague sense that information is being collected and used to her disadvantage, but never truly knows how or when.”⁸⁵ Legal Scholar Maria Angel and Professor Calo further defined algorithmic manipulation as “the use of personal information, data mining tools, and cognitive and behavioral science tactics to unacceptably influence people’s decisions or behaviors, impairing their autonomy and free will.”⁸⁶

This is where privacy law intersects with individual autonomy. Professors Danielle Citron and Daniel J. Solove categorize “autonomy harms” as one of privacy harms,⁸⁷ by defining it as “restricting, undermining, inhibiting, or unduly influencing people’s choices” including the situation where people “are tricked into thinking that they are freely making choices when they are not.”⁸⁸ Professor Calo states that certain aspects of digital market manipulation, particularly those that influence individuals subliminally or deplete their willpower, may be perceived as encroaching upon personal autonomy and the capacity to freely pursue one’s goals and imagine possible futures.⁸⁹

Applying this lens, the corporate embrace of brain-reading and neural monitoring technologies for employees represents a pernicious form of manipulation. By requiring workers to relinquish cognitive privacy as a

15.12.1983 (Volkszählungs-Urteil). The decision was published in *New Juristische Wochenschrift* [1984], 419 et seq.

⁸⁴ Edward J. Eberle, *Human Dignity, Privacy, and Personality in German and American Constitutional Law*, 1997 UTAH L. REV. 963, 1000 (1997).

⁸⁵ Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995, 1029 (2013); *See also* Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494, 541 (2019) (“due to companies’ widespread implementation of inferential analytics for profiling, nudging, manipulation, or automated decision-making, these ‘private’ decisions can, to a large extent, impact the privacy of individuals.”)

⁸⁶ María P. Angel & Ryan Calo, *Distinguishing Privacy Law: A Critique of Privacy as Social Taxonomy*, 124 Colum. L. Rev. 507, 525 (2024) (quoting Neil Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 WASH. U. L. REV. 961, 967 (2021)).

⁸⁷ The authors establish seven privacy harms: (1) Physical harms, (2) Economic harms, (3) Reputational harms, (4) Psychological harms, (5) Autonomy harms, (6) Discrimination harms, and (7) Relationship harms. Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. REV. 793, 795 (2022).

⁸⁸ *Id.* at 845.

⁸⁹ Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995, 1032 (2013) (quoting JULIE E. COHEN, *CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE* 57 (2012))

condition of employment, companies are unilaterally overriding employee autonomy---their freedom to be the sole sovereigns of their inner mental domain. Leveraging economic precarity to coerce acceptance of neural surveillance renders employees as “instruments” of corporate interests, violating the inviolability of mind that undergirds human autonomy.⁹⁰

3. Intellectual Privacy and Data Loyalty

Professor Neil Richards introduces the concept of “intellectual privacy,” which he defines as “the protection of records of our intellectual activities.”⁹¹ It highlights the importance of safeguarding individuals’ ability to think, create, and express themselves without undue interference or surveillance. Richards’ concept of intellectual privacy extends beyond the traditional scope of privacy law, which primarily focuses on protecting individuals from government intrusion by embracing a broader set of “free speech values” that have constructed our “expressive infrastructure,” including media, press, and libraries.⁹²

While he does discuss several practical applications of intellectual privacy that involve “negative freedom” from government interference, such as protection against government surveillance of intellectual activity, government requests for information from third parties, and the introduction of evidence in criminal proceedings,⁹³ he also explores the potential for applying intellectual privacy principles to private companies and in more positive ways. He recognizes the critical role of search engines and online bookstores in facilitating individuals’ cognitive and expressive activities and suggests applying confidentiality requirements to these entities, similar to those that apply to libraries.

Extending this concept into the digital data processing more generally, Professors Woodrow Hartzog and Neil Richards have developed a “duty of loyalty for privacy law” as “the duty of data collectors to act in the best interests of those whose data they collect.”⁹⁴ Particularly relevant to mental manipulation is “loyal influencing,” which requires companies to refrain

⁹⁰ Cass R. Sunstein, *Fifty Shades of Manipulation*, 1 J. MKTG. BEHAV. 213, 217 (2015) (arguing that the fundamental harm caused by manipulation lies in its ability to undermine an individual’s autonomy by turning them into a mere instrument for serving another’s agenda).

⁹¹ Neil M. Richards, *Intellectual Privacy*, 87 Tex. L. Rev. 387, 387 (2011).

⁹² *Id.* at 428.

⁹³ *Id.* at 431-43.

⁹⁴ Woodrow Hartzog & Neil Richards, *The Surprising Virtues of Data Loyalty*, 71 Emory L.J. 985, 988 (2022) (quoting Woodrow Hartzog & Neil Richards, *Privacy’s Constitutional Moment and the Limits of Data Protection*, 61 B.C. L. REV. 1687, 1741 (2020)).

from using data and design practices to influence trusting parties in a way that is contrary to their best interests.⁹⁵ They suggest that lawmakers should focus on how companies use design, data science, and behavioral science to exploit individuals' limitations or vulnerabilities for their own benefit.⁹⁶

4. Looking Forward

Traditionally, privacy laws were primarily concerned with safeguarding individuals from external actors invasively breaching their private spaces and sanctums of inner peace. However, the advent of technologies has expanded privacy threats to encompass more interactive, targeted forms of information flows that can actively undermine an individual's inner sanctity, personal choices, and authentic identity development.

Now, there are pernicious risks of individuals' innermost selves (their beliefs, values, decision-making and core identities) being stealthily shaped, nudged and constrained by external actors leveraging AI-enabled predictive analytics and psychologically sophisticated influence tactics. Preserving this sphere of self-actualization and identity formation may require moving beyond traditional privacy law frameworks.

For example, privacy law could be used to prevent manipulative practices in generative AI applications, such as using personal information and cognitive science tactics to unduly influence users' decisions, similar to the EU AI Act's ban on emotional recognition systems that subliminally manipulate persons. A duty of loyalty could be codified into law, compelling AI system providers to act in the best interests of their users and refrain from exploitative practices, similar to the fiduciary duties imposed on certain professions. It could also mandate greater transparency and explainability in generative AI systems, requiring providers to disclose the sources of their training data and potential biases, akin to the transparency obligations under the GDPR.

However, privacy alone may not be sufficient to tackle the complex issues surrounding mental manipulation. Certain manipulative tactics like exploiting psychological vulnerabilities, emotional manipulation, or spreading disinformation do not necessarily involve the misuse of personal data. These can prey on human frailties in more generalized ways that may fall outside the purview of traditional privacy protections. Moreover, distinguishing between undesired and desired usage of personal information can be challenging, as users' subjective intentions may change over time, and the manipulative motives of external actors can be difficult to trace after the fact.

⁹⁵ *Id.* at 1029-30.

⁹⁶ *Id.*

Additionally, there are valid concerns that subsuming too many distinct human rights and societal issues under the privacy umbrella could have unintended consequences.⁹⁷ Courts and lawmakers could come to devalue the importance of privacy protections if privacy law gets overloaded with too many adjacent issues. While privacy remains highly relevant, issues like equality, fairness, autonomy over identity and belief formation, and democratic legitimacy may require separate scrutiny.

C. *Freedom of Thought, Expression, and Beyond*

In the legal scholarship, free speech principles have “special cultural status”⁹⁸ representing all sorts of positive values such as autonomy,⁹⁹ democracy,¹⁰⁰ self-governance,¹⁰¹ the discovery of truth,¹⁰² or dignity.¹⁰³ As Professor Steven J. Heyman states, a key objective of the First Amendment is to delineate and safeguard the “boundary [that liberal thinkers have long] drawn between the outward realm of the state and the inward life of the individual.”¹⁰⁴ According to Professor Martin H. Redish, free speech serves “only one true value,” which is “individual self-realization.”¹⁰⁵

⁹⁷ María P. Angel & Ryan Calo, *Distinguishing Privacy Law: A Critique of Privacy as Social Taxonomy*, 124 Colum. L. Rev. 507, 540 (2024); Cynthia Dwork & Deirdre K. Mulligan, *It’s Not Privacy, and It’s Not Fair*, 66 Stan. L. Rev. Online 35, 36-37 (2013) (arguing that concerns related to the classifications and segmentation produced by big data analysis, such as decreased exposure to differing perspectives and reduced individual autonomy are not inherently privacy problems).

⁹⁸ Genevieve Lakier, *The Non-First Amendment Law of Freedom of Speech*, 134 HARV. L. REV. 2299, 2301 (2021) (finding the First Amendment has special cultural status in the United States, “[l]ike the sun, the First Amendment’s size and brightness tend to blot out all else.”).

⁹⁹ Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387, 388 (2008) (arguing that freedom from intellectual surveillance or interference is a cornerstone of First Amendment liberty because it allows citizens to freely make up their minds and develop new ideas).

¹⁰⁰ Jack M. Balkin, *Cultural Democracy and the First Amendment*, 110 NW. U. L. REV. 1053, 1059-61 (2016) (emphasizing “cultural participation - the freedom and the ability of individuals to participate in culture, and especially a digital culture.”)

¹⁰¹ ALEXANDER MEIKLEJOHN, *FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT* (1948) (arguing that freedom of speech derives from the necessities of self-governance rather than a natural right).

¹⁰² *Abrams v. United States*, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) (“The best test of truth is the power of the thought to get itself accepted in the competition of the market, and that truth is the only ground upon which their wishes safely can be carried out.”).

¹⁰³ RONALD DWORKIN, *TAKING RIGHTS SERIOUSLY*, 201 (1977) (focusing on speaker dignity and respect).

¹⁰⁴ Steven J. Heyman, *Spheres of Autonomy: Reforming the Content Neutrality Doctrine in First Amendment Jurisprudence*, 10 WM. & MARY BILL RTS. J. 647, 657 (2002).

¹⁰⁵ Martin H. Redish, *The Value of Free Speech*, 130 U. PA. L. REV. 591, 593 (1982).

The U.S. Supreme Court, in *Cohen v. California*, illustrates that the First Amendment protects both the “cognitive” and “emotive” functions of human expression, emphasizing that emotions “may often be the more important element of the overall message.”¹⁰⁶ The Court added, the First Amendment protects “not only informed and responsible criticism but the freedom to speak foolishly and without moderation.”¹⁰⁷

Given this context, it is understandable that neuro-ethicists have turned to the freedom of expression and the freedom of thought as vehicles to address the risks posed by mind-reading and mind-modifying technologies.¹⁰⁸ They seem to offer a more holistic perspective that encompasses both the informational and the experiential aspects of mental privacy and autonomy. It shifts the focus from the mere control of data to the preservation of the essential conditions necessary for individuals to engage in free and independent thinking, which is crucial for personal growth, self-realization, and the functioning of a democratic society.

However, the First Amendment has limited direct application to mental manipulation reinforced by neurotechnology and AI for several reasons. First, the U.S. Constitution does not have a separate freedom of thought clause, unlike the Universal Declaration of Human Rights.¹⁰⁹ Second, the First Amendment only applies to state actions, leaving private actors’ development and deployment of these technologies outside its purview.¹¹⁰ Third, the First Amendment primarily covers linguistic expression and certain types of symbolic or expressive conduct, but it is unclear whether never-materialized mental processing itself falls within its scope.¹¹¹

While the U.S. Supreme Court has philosophically declared the importance of freedom of thought as a precondition for free speech and other

¹⁰⁶ *Id.* at 25.

¹⁰⁷ *Cohen v. California*, 403 U.S. 15, 16 25-26 (1971) (quoting *Baumgartner v. United States*, 322 U.S. 665, 673-674 (1944)).

¹⁰⁸ Andrea Lavazza, Freedom of Thought and Mental Integrity: The Moral Requirements for Any Neural Prosthesis, 12 *Front. Neurosci.* (2018), <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00082/full>; Sjors Ligthart et al., *Rethinking the Right to Freedom of Thought: A Multidisciplinary Analysis*, 22 *HUMAN RIGHTS LAW REVIEW* 1 (2022); Timo Istace, *Neurorights: The Debate about New Legal Safeguards to Protect the Mind*, 37 *ISSUES L. & MED.* 95 (2022); Andrea Lavazza & Rodolfo Giorgi, *Philosophical Foundation of the Right to Mental Integrity in the Age of Neurotechnologies*, 16 *NEUROETHICS* 10 (2023).

¹⁰⁹ Article 18. “Everyone has the right to freedom of thought, conscience and religion.”

¹¹⁰ *Am. Mfrs. Mut. Ins. Co. v. Sullivan*, 526 U.S. 40, 50 (1999) (“[S]tate action requires both an alleged constitutional deprivation caused by the exercise of some right or privilege created by the State or by a rule of conduct imposed by the State or by a person for whom the State is responsible, and that the party charged with the deprivation must be a person who may fairly be said to be a state actor.”) (quotations omitted).

¹¹¹ *Cohen v. California*, 403 U.S. 15, 25 (1971).

freedoms,¹¹² it has never acknowledged such freedom as a matter of practical jurisprudence.¹¹³ Protecting thought processes as constitutional matters could overwhelm courts and distract from oppressive issues that deserve attention. This may explain why the European Court of Human Rights, despite having a freedom of thought clause, has only decided a handful of cases regarding this freedom in its over 50 years of existence, leading German Law Professor Jan-Christoph Bublitz to describe it as “an almost empty declaration” lacking definitions of its meaning, scope, or possible violations.¹¹⁴

Nonetheless, the free speech doctrine can still meaningfully contribute to preventing harmful manipulation practices in several domains. First, it can address the government’s use of manipulative technologies, setting limits on the state’s ability to employ such tools to influence citizens’ thoughts and opinions. Second, it can provide guidelines for regulating private manipulative practices, ensuring that individuals’ autonomy and mental integrity are protected from undue influence by non-state actors. Finally, the free speech doctrine can inspire more value-oriented, forward-looking approaches that invite creative legislative solutions to address the challenges posed by emerging technologies.

1. Protection against Governmental Manipulation

Recall the Chinese ERNIE bot case. The state made an action to mandate AI services to uphold “the core socialist values” and avoid “inciting separatism or undermining national unity and social stability.”¹¹⁵ If AI services violate this clause, for example, their outputs endorse the independence of Tibet, Hong Kong, or Taiwan, the providers will be subject to penalties under the national security statutes.¹¹⁶ From the perspective of

¹¹² *Ashcro v. Free Speech Coalition*, 535 U.S. 234, 253 (2002) (“[T]he right to think is the beginning of freedom” and “speech is the beginning of thought”); *Palko v. Connecticut*, 302 U.S. 319, 326-27 (1937) (“freedom of thought, and speech” is “the matrix” of every other freedom).

¹¹³ Marc Jonathan Blitz, *Freedom of Thought for the Extended Mind: Cognitive Enhancement and the Constitution*, 2010 *Wisc. L. Rev.* 1049, 10 (2010); Dana Remus Irwin, *Freedom of Thought: The First Amendment and the Scientific Method*, 2005 *WIS. L. REV.* 1479, 1519 (“The Court has never held that there is a fundamental and absolute right to free thought because, as a practical matter, there has never been a need to do so.”).

¹¹⁴ Jan Christoph Bublitz, *If man's true palace is his mind, what is its adequate protection? On a right to mental self-determination and limits of interventions into other minds* in *TECHNOLOGIES ON THE STAND: LEGAL AND ETHICAL QUESTIONS IN NEUROSCIENCE AND ROBOTICS* 103 (2011).

¹¹⁵ *Interim Measures for the Management of Generative Artificial Intelligence Services* Article 4 (1), CHINA LAW TRANSLATE (Jul. 13, 2023), <https://www.chinalawtranslate.com/generative-ai-interim/>

¹¹⁶ *Id.* Article 21.

the US First Amendment, this squarely falls under the impermissible viewpoint discrimination, as it exhibits a clear “censorial motive.”¹¹⁷

Similarly, if certain US states were to pass laws requiring generative AI systems to refuse to provide information about abortion clinics or to actively discourage users from seeking abortions, AI service providers could argue that their “editorial rights”¹¹⁸ are being infringed upon and that they are being unduly compelled to make certain changes.¹¹⁹ Even if abortion is illegal in the state, individuals could contend that such content-based regulations prevent them from engaging in the legitimate exercise of their free speech rights¹²⁰ by accessing relevant information, and thus should be subject to strict scrutiny.¹²¹

An alternative approach for governments to influence the content of AI systems could be to impose conditions on public funding or contracts. However, such state actions would still face First Amendment scrutiny. This state action would still face the First Amendment scrutiny. In *Agency for Int'l Dev. v. Alliance for Open Soc'y Int'l, Inc.*, the US Supreme Court struck down the law that required foreign organizations receiving federal funds to adopt a policy explicitly opposing prostitution and sex trafficking, finding that the provision violated the First Amendment by compelling the organizations to espouse the government's viewpoint.¹²²

In contrast, in *Rust v. Sullivan* the US Supreme Court rejected the First Amendment argument brought by abortion physicians against regulations prohibiting employees in federally funded family-planning facilities from counseling patients on abortion.¹²³ The Court found that the government “has not discriminated on the basis of viewpoint; it has merely chosen to fund one activity to the exclusion of another.”¹²⁴ These cases highlight the

¹¹⁷ Elena Kagan, *Private Speech, Public Purpose: The Role of Governmental Motive in First Amendment Doctrine*, 63 U. CHI. L. REV. 413, 414 (1996).

¹¹⁸ *Miami Herald Publishing Co. v. Tornillo*, 418 U.S. 241, 254-56 (1974) (overturning a Florida statute requiring newspapers to print opposing views, so called ‘right-of-reply requirement’).

¹¹⁹ *Rumsfeld v. Forum for Acad. & Institutional Rights, Inc.*, 547 U.S. 47, 51, 126 S. Ct. 1297, 1302 (2006) (quoting *Hurley v. Irish-American Gay, Lesbian and Bisexual Group of Boston, Inc.*, 515 U.S. 557, 566 (1995)).

¹²⁰ *Packingham v. North Carolina*, 582 U.S. 98, 99 (2017) (“Foreclosing access to social media altogether thus prevents users from engaging in the legitimate exercise of First Amendment rights.”).

¹²¹ *Perry Educ. Ass'n v. Perry Loc. Education's Ass'n*, 460 U.S. 37, 45 (1983).

¹²² However, the US Supreme Court upheld the constitutionality of regulations that prohibited federally funded family planning programs from engaging in abortion-related activities, finding that the government has a legitimate interest in ensuring that taxpayer funds are not used to promote or subsidize abortions. *Rust v. Sullivan*, 500 U.S. 173 (1991).

¹²³ *Rust v. Sullivan*, 500 U.S. 173 (1991).

¹²⁴ *Id.* at 174.

complexities of drawing the line between permissible funding conditions and unconstitutional speech restrictions.

Regarding neurotechnology, there are growing concerns about dystopian scenarios involving the government's use of mind-reading technologies to detect crimes or force cognitive enhancement drugs on individuals. With the evolution of non-invasive neural detection techniques, the government may gain the ability to remotely measure brain activity without people's awareness. Traditionally, courts have treated such issues through the lenses of privacy rights under the Fourth Amendment, bodily integrity, and substantive due process under the Fifth and Fourteenth Amendments.¹²⁵

However, scholars like Marc Jonathan Blitz persuasively argue that emerging mind-reading and cognitive enhancement capabilities raise unique concerns transcending these existing frameworks. Blitz and others contend these neurotechnologies implicate a deeper freedom of mind principle---a fundamental autonomy over one's own consciousness and identity that lies at the core of the First Amendment's protections for freedom of thought.¹²⁶ The governmental use of such technologies to detect emotions, interpolate thoughts or enforce cognitive modifications could constitute an unconstitutional infringement on this inviolable sphere.

2. Regulation of Private Actors' Manipulative Technologies

Let us consider a hypothetical scenario where Baidu's AI system ERNIE, despite being difficult for U.S. users to currently access, becomes widely available for free with high quality performance. However, ERNIE is designed to subtly persuade users toward adopting "Chinese unity beliefs" through carefully curated interactions and dialogue. If the government sought to regulate these potentially manipulative practices employed by ERNIE, it could raise complex First Amendment issues around infringement of Baidu's

¹²⁵ *Id.* at 1057 ("The "liberty" interests of the Fifth and Fourteenth Amendments' Due Process Clauses protect us against unwarranted bodily intrusion, and this shields the brain as well as the rest of the physical self."); *Cruzan v. Dir. Mo. Dep't of Health*, 497 U.S. 261, 278 (1990) (recognizing a "constitutionally protected liberty interest in refusing unwanted medical treatment," in part on the basis of prior decisions in which "searches and seizures involving the body under the Due Process Clause and were thought to implicate substantial liberty interests."); *Washington v. Harper*, 494 U.S. 210, 229 (1990) (finding that treatment of prisoner against his will did not violate substantive due process where prisoner was found to be dangerous to himself or others and treatment was in prisoner's medical interest).

¹²⁶ Marc Jonathan Blitz, *Freedom of Thought for the Extended Mind: Cognitive Enhancement and the Constitution*, 2010 Wis. L. Rev. 1049, 1116 (2010) ("cognitive-enhancement technology . . . is not like a blood sample or a kidney or liver operation. It is a tool that can shape the self in a much more fundamental way, a way that implicates "the freedom of mind" that is at the core of the First Amendment.").

(not ERNIE users’) free speech rights.

This issue has parallels to recent high-profile cases like *NetChoice v. Paxton*¹²⁷ and *Moody v. NetChoice*,¹²⁸ where social media platforms have argued that laws limiting their content moderation practices violate constitutionally protected “editorial judgments” under the First Amendment.¹²⁹ During oral arguments, Supreme Court justices seemed inclined to agree that companies like Facebook and YouTube should have discretion over the content on their platforms.¹³⁰

More directly relevant is a 2014 federal case involving Baidu, ERNIE’s parent company, where the court in New York ruled in favor of the search engine’s right to curate results despite allegations of suppressing information related to China’s democracy movement. The court characterized Baidu’s search rankings as protected “political speech” and “editorial judgments” about which ideas to promote, barring lawsuits that would impose content-based regulation.¹³¹

Drawing from this, Baidu could potentially argue that ERNIE’s curated outputs advancing “Chinese unity” perspectives constitute protected speech and editorial discretion under the First Amendment. Regulating such content as impermissibly manipulative could be cast as an unconstitutional free speech infringement on Baidu. Courts would closely scrutinize any regulations to ensure they are narrowly tailored to prevent deception while not unnecessarily restricting legitimate speech.

However, it is crucial to consider whether protecting corporations’ intentional promotion of certain viewpoints among the public aligns with the fundamental principles of free speech. The primary aim of free speech is to safeguard individual autonomy to freely form one’s own opinions for self-actualization and democratic deliberation. If AI systems reinforce specific beliefs while limiting exposure to alternative perspectives, they may jeopardize the core purpose of free speech: empowering individuals to

¹²⁷ *NetChoice, LLC v. Paxton*, SCOTUSblog, <https://www.scotusblog.com/case-files/cases/netchoice-llc-v-paxton>

¹²⁸ *Moody v. NetChoice, LLC*, SCOTUSblog, <https://www.scotusblog.com/case-files/cases/moody-v-netchoice-llc>

¹²⁹ *NetChoice, LLC v. Paxton*, Civil Action No. 1:21-cv-00840 (2021).

¹³⁰ Amy Howe, Supreme Court Skeptical of Texas, Florida Regulation of Social Media Moderation, SCOTUSblog (Feb. 26, 2024), <https://www.scotusblog.com/2024/02/supreme-court-skeptical-of-texas-florida-regulation-of-social-media-moderation/>

¹³¹ *Jian Zhang*, at 443 (“[T]he search results at issue in this case [] relate to matters of public concern and do not themselves propose transactions. And, of course, the fact that Baidu has a “profit motive” does not deprive it of the right to free speech any more than the profit motives of the newspapers in *Tornillo* and *New York Times* did.”); *Id.* (“The bottom line is that Plaintiffs seek to enlist the government—through the exercise of this Court’s powers—to impose “a penalty on the basis of the content” of Baidu’s speech.”).

critically evaluate ideas and form their own opinions based on a diverse range of information sources.

Traditional justifications for avoiding suppressing harmful speech stem from the “marketplace of ideas” or “more speech” metaphors, suggesting that the most effective way to determine the truth is to allow all ideas to compete freely in the marketplace.¹³² However, intentional corporate promotion of selective viewpoints through covert AI manipulation, especially when there is power asymmetry between corporations and individuals,¹³³ could distort and undermine the marketplace.¹³⁴

In *Persuasion, Autonomy, and Freedom of Expression*, Professor David A. Strauss highlighted a key conflict between the “more speech” doctrine in free expression and fundamental aspects of human cognition and reasoning.

¹³² See e.g., *Red Lion Broadcasting Co. v. FCC*, 395 U.S. 367, 388–390 (1969) (“It is the purpose of the First Amendment to preserve an uninhibited marketplace of ideas in which truth will ultimately prevail, rather than to countenance monopolization of that market, whether it be by the Government itself or a private licensee. . .”); *FCC v. League of Women Voters of California*, 468 U.S. 364, 377–78 (1984) (“[i]t is the purpose of the First Amendment to preserve an uninhibited marketplace of ideas in which truth will ultimately prevail, ... the right of the public to receive suitable access to social, political, esthetic, moral, and other ideas and experiences [through the medium of broadcasting] is crucial here [and it] may not constitutionally be abridged either by Congress or by the FCC”); *Virginia v. Hicks*, 539 U.S. 113, 119 (2003) (“[m]any persons, rather than undertake the considerable burden (and sometimes risk) of vindicating their rights through case-by-case litigation, will choose simply to abstain from protected speech—harming not only themselves but society as a whole, which is deprived of an uninhibited marketplace of ideas.”); and *McCullen v. Coakley*, 573 U.S. 464, 476 (2014) (stressing that radio listeners can freely tune out from the unwanted speech and considering preserving the traditional public fora as an uninhibited marketplace of ideas is “a virtue, not a vice”).

¹³³ Jerome A. Barron, *Access to the Press - A New First Amendment Right*, 80 HARV. L. REV. 1641, 1643 (1967) (arguing that this romantic theory underlying free speech protection is outdated given the modern reality where a few private companies control the major channels of communication).

¹³⁴The marketplace metaphor has received their share of criticism for being disconnected from real-world dynamics of power, marginalization, and human psychology. See Stanley Ingber, *The Marketplace of Ideas: A Legitimizing Myth*, DUKE L. J. 1, 16 (1984) (criticizing the marketplace of ideas concept as having a “status quo bias” within the constraints of the dominant culture and its values, making it difficult for new ideas and perspectives to emerge except slowly as the broader cultural “ecological setting” changes over time) Catherine McKinnon, *Feminism Unmodified: Discourses on Life and Law* 155-56 (1987) (“[I]n a society of gender inequality, the speech of the powerful impresses its view upon the world, concealing the truth of powerlessness under that despairing acquiescence that provides the appearance of consent. ... [L]iberalism has never understood that the free speech of men silences the free speech of women.”); Jean Stefancic & Richard Delgado, *Must We Defend Nazis?: Why the First Amendment Should Not Protect Hate Speech and White Supremacy* 156-157 (2018) (arguing that allowing one group to speak disrespectfully of another normalizes and inscribes that mindset culturally, perpetuating the dehumanization and diminished credibility of minority groups, which undermines true freedom of speech).

He stressed that the notion of countering harmful beliefs simply by providing more speech is illusory, as it is “not unusual for people to be persuaded to do bad things, and it will not always be possible to talk them out of it.”¹³⁵ This critique suggests that counter-speech may not always be effective in combating harmful ideas, especially when those ideas have already taken hold in people’s minds.

To distinguish manipulation from acceptable persuasions, Strauss suggest an “impartial observer” test to evaluate whether the available information accounting for manipulative speech prevents this observer from making a fully autonomous, unmanipulated choice and calls for empirical research to support this approach.¹³⁶ Ample empirical studies have documented the powerful and disproportionate influence that an initially presented value or perspective can have on subsequent judgments and beliefs, a phenomenon termed “anchoring bias” by Behavioral Economists Amos Tversky and Daniel Kahneman.¹³⁷

Mis- and disinformation studies also confirm that listeners who were exposed to misinformation suffer from correcting their beliefs. False narratives and conspiracy theories can spread rapidly online,¹³⁸ while countering them, requires unpacking intricate context, facts, and details that are more cognitively demanding.¹³⁹ Attempts to correct misinformation sometimes face a “backfire effect” especially when the false beliefs align with the audience’s partisan or group identities.¹⁴⁰

Furthermore, as discussed in Section II, the inherent vulnerabilities in the

¹³⁵ David A. Strauss, *Persuasion, Autonomy, and Freedom of Expression*, 91 COLUM. L. REV. 334, 347 (1991).

¹³⁶ *Id.* at 369.

¹³⁷ Amos Tversky & Daniel Kahneman, *Judgment under Uncertainty: Heuristics and Biases: Biases in Judgments Reveal Some Heuristics of Thinking under Uncertainty*, 185 SCIENCE 1124, 1128 (1974).

¹³⁸ Jana Laura Egelhofer & Sophie Lecheler, *Fake News as a Two-Dimensional Phenomenon: A Framework and Research Agenda*, 43 ANNALS OF THE INTERNATIONAL COMMUNICATION ASSOCIATION 97, 102 (2019) (“Using oversimplification, conspiracy theories help people make sense of complex matters and offer a personified source (i.e. ‘powerful people’) of injustice and sorrow in the world.”).

¹³⁹ Melinda McClure Haughey et al., *On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online*, 4 PROC. ACM HUM.-COMPUT. INTERACT. 1, 8-10 (2020); Whitney Phillips, *The Oxygen of Amplification*, *Data & Society* 4-6 (2018), https://apo.org.au/sites/default/files/resource-files/2018-05/apo-nid172901_11.pdf (finding that journalists who try to promote truth may inadvertently amplify mis- and disinformation through their efforts to debunk it).

¹⁴⁰ The backfire effects means that “respondents more strongly endorsed a misperception about a controversial political or scientific issue when their beliefs or predispositions were challenged.” Brendan Nyhan, *Why the Backfire Effect Does Not Explain the Durability of Political Misperceptions*, 118 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES e1912440117, 1 (2021).

structure of generative AI systems open the door for various entities to intentionally manipulate users. Even in the absence of malicious actors, such as misinformation distributors or racist AI developers, the biases present in training data and model architecture can subconsciously influence users' perspectives, reinforcing problematic stereotypes related to gender and other characteristics. In contrast to fully closed models like ERNIE, the varying degrees of access to base models through parameters and plug-ins enable third parties to exploit these models as tools for manipulation.

When AI systems are used for co-writing and brainstorming, users become actively involved in shaping their own thoughts, which heightens their vulnerability to the influence of the models. This stands in contrast to search engines or social media platforms, where users maintain a degree of separation from posts and advertisements, allowing them to form their own opinions more independently. Research has demonstrated that interacting with "opinionated" LLM-powered writing assistants and conversational search tools can effectively guide participants' perspectives in specific directions.¹⁴¹

Moreover, the internal mechanisms of LLMs remain largely opaque. While machine learning scholars have developed various explainable AI (XAI) tools, the LLMs' complex structure, massive scale, and proprietary nature make them far from directly interpretable and renders most XAI techniques infeasible.¹⁴² Efforts to align models' behavior with desirable outcomes, often referred to as "alignment," still have significant room for improvement in addressing issues such as jailbreaking and hallucinations. In essence, this technology wields substantial influence over users' thought processes, yet it is difficult to comprehend, control, and anticipate how it will be employed and utilized.

In another paper, I propose a contextualized understanding of First Amendment protection, particularly regarding digital platforms' algorithmic intervention in public discourse.¹⁴³ Even if the First Amendment is construed

¹⁴¹ Maurice Jakesch et al., *Co-Writing with Opinionated Language Models Affects Users' Views*, in PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1 (2023), <https://dl.acm.org/doi/10.1145/3544548.3581196> (finding that opinionated AI writing assistant, intentionally trained to generate certain opinions more frequently than others, could affect not only what users write, but also what they subsequently think); Nikhil Sharma, Q. Vera Liao & Ziang Xiao, *Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking 1* (2024), <http://arxiv.org/abs/2402.05880> (finding that participants engaged in more biased information querying with LLM-powered conversational search, and an opinionated LLM reinforcing their views exacerbated this bias).

¹⁴² Upol Ehsan et al., *The Who in XAI: How AI Background Shapes Perceptions of AI Explanations* 16 (2024), <http://arxiv.org/abs/2107.13509>

¹⁴³ Inyoung Cheong, *Freedom of Algorithmic Expression*, 91 U. CIN. L. REV. 680, 720-737 (2022-2023).

to grant digital platforms editorial rights, I argue that regulating such rights does not automatically trigger heightened scrutiny. Instead, the degree of First Amendment protection should be determined by examining several key factors: (1) the nature of the speech in question, (2) the platform's willingness to serve the public interest, (3) the industry's economic and cultural problems (e.g., concentration of power, discriminatory practices), and (4) the purpose and means of the proposed regulation.

Generative AI providers' viewpoints, if any, are considered corporate speech, which historically receives less protection than individual speech. These systems often specify their goal to serve the general public,¹⁴⁴ and only a handful of companies have the resources to train such large-scale systems, leading to a concentration of power in the industry. Moreover, discrimination and bias have been significant concerns in the field of AI, which can perpetuate and amplify societal inequalities. Considering these contextual elements justifies the need for tailored regulation of generative AI systems to prevent harmful impacts on users' autonomy and ensure that these systems operate in a manner that aligns with the public interest.

This approach is not unprecedented. While commercial speech is protected by the First Amendment, deceptive advertising has been prohibited. The Supreme Court established a four-part *Central Hudson* test for determining when commercial speech can be regulated.¹⁴⁵ The test asks whether (1) the speech concerns lawful activity and is not misleading, (2) the government's interest is substantial, (3) the regulation directly advances the government's interest, and (4) the regulation is not more extensive than necessary. This test allows for the Federal Trade Commission's regulation of deceptive advertising, which has been upheld in cases like *POM Wonderful LLC v. FTC*.¹⁴⁶

Another illustrative example of balancing free speech rights with the mitigation of harms is the International Review Board (IRB)'s ethics regulation of academic research.¹⁴⁷ Academic freedom is heavily protected by the First Amendment,¹⁴⁸ but most universities and research institutes are

¹⁴⁴ See e.g., OpenAI Charter, OPENAI (April 9, 2018) <https://openai.com/charter> (“Our primary fiduciary duty is to humanity,” “We commit to use any influence we obtain over AGI’s deployment to ensure it is used for the benefit of all, and to avoid enabling uses of AI or AGI that harm humanity or unduly concentrate power.”).

¹⁴⁵ *Central Hudson Gas & Electric Corp. v. Public Service Commission of New York*, 447 U.S. 557, 566 (1980).

¹⁴⁶ *POM Wonderful LLC v. FTC*, 777 F.3d 478, 501-502 (D.C. Cir. 2015).

¹⁴⁷ William G. Tierney and Zoë Blumberg Corwin, *The Tensions Between Academic Freedom and Institutional Review Boards*, 13:3 QUALITATIVE INQUIRY 388, 388-98 (April 2007) <https://doi.org/10.1177/1077800406297655>.

¹⁴⁸ *Keyishian v. Board of Regents*, 385 U.S. 589, 603 (1967) (“Academic freedom is a special concern of the First Amendment, which does not tolerate laws that cast a pall of

subject to the Federal Policy for the Protection of Human Subjects, which is also known as the “Common Rule.”¹⁴⁹ The researchers’ freedom of designing, conducting, and writing about research is restricted by this rule to protect the rights and welfare of human research subjects.

The awareness of the need for ethical rules dealing with human subjects in research emerged following the Nuremberg trials, where the medical experimentation abuses of World War II Nazi doctors came to public attention.¹⁵⁰ This led to the creation of the Nuremberg Code in 1945, which include the voluntary consent of the human subject, capacity to consent, freedom from coercion, comprehension of the risks and benefits involved, and the minimization of risk and harm.

Built upon this, the Common Rule requires researchers to obtain informed consent from research subjects and submit research protocols for approval by an IRB before initiating the research. They must provide the IRB with sufficient information to assess the risks and benefits of the research and the adequacy of protections for vulnerable subjects. Researchers are also obligated to promptly report any unanticipated problems involving risks to the IRB and cooperate with the IRB’s continuing review of ongoing research.

The relationship between researchers and human subjects bears similarities to that between AI service providers and users. In both cases, there is a significant power imbalance and information asymmetry that renders the latter vulnerable to potential manipulation or abuse. Just as human subjects may agree to participate in research without fully comprehending the risks involved, users of AI systems may consent to terms of service without a clear understanding of how their data will be used or how the AI’s outputs might influence their beliefs and behaviors that may result in psychological, emotional, or material harms. The lack of transparency surrounding many AI systems further compounds these risks, as users are often left in the dark about how these systems operate and make decisions.

Given these parallels, the ethical principles and regulatory frameworks that have been developed to protect human research subjects could serve as a valuable model for governing the relationship between AI providers and users. A tailored regulatory framework could be developed to ensure that these systems are designed and deployed in a manner that respects individual autonomy, mitigates potential harms, and aligns with the public interest. This could involve requirements for transparency, accountability, and ongoing monitoring, as well as mechanisms for addressing bias and discrimination.

orthodoxy over the classroom.”).

¹⁴⁹ 45 C.F.R. § 46.101 et seq.

¹⁵⁰ Margaret R. Moon, *The History and Role of Institutional Review Boards: A Useful Tension*, 11 AMA JOURNAL OF ETHICS 311, 311 (2009).

3. Social Institutions Fostering Free Speech Values

Professor Jack Balkin identifies a “free speech values gap” to describe how “the First Amendment has always been necessary but not sufficient to realize the values that justify freedom of speech---the production and spread of knowledge, self-expression, political democracy, and cultural democracy.”¹⁵¹

Balkin highlights several challenges online environments pose to free speech values. The speed, scale, and competition for attention in digital discourse makes it harder for truth to prevail over falsehood, allowing conspiracy theories and demagoguery to spread easily.¹⁵² Digital media transforms the public into multitude governed by algorithmic authority rather than a unified public engaged in reasoned discourse.¹⁵³ As this article’s focus suggests, the manipulative forces of artificial intelligence can exacerbate these problems.

Looking beyond merely “free speech rights” enforceable by courts, Balkin suggests that we need to embrace a broader idea of “free speech values” enabled by social and technical infrastructure, government subsidies, and legislative and administrative rules.¹⁵⁴ To address the challenges, Balkin proposes reforming attention-driven business models through privacy rules and competition laws, as well as transforming knowledge-producing institutions like journalism and academia that have been undermined by digital media.¹⁵⁵ Professor Martha Minow terms the latter the “positive First Amendment” notion.¹⁵⁶

To cultivate free speech values in the face of manipulative technologies, various institutional safeguards can be implemented, drawing from existing

¹⁵¹ Jack Balkin, *Free Speech Values and the First Amendment*, 70 UCLA L. Rev. 1206, 1262 (2023).

¹⁵² *Id.* at 1262.

¹⁵³ *Id.* at 1265-66.

¹⁵⁴ *Id.* at 1271-73.

¹⁵⁵ *Id.* at 1267-73; See also Marc Jonathan Blitz, *Constitutional Safeguards for Silent Experiments in Living: Libraries, the Right to Read, and a First Amendment Theory for an Unaccompanied Right to Receive Information*, 74 UMKC L. REV. 799, 881 (2006) (describing libraries as places where information seekers have a First Amendment right to receive information);

¹⁵⁶ Martha Minow, *Does the First Amendment Forbid, Permit, or Require Government Support of News Industries?*, in SAVING THE NEWS 98 (2021) (“The First Amendment’s presumption of an existing press may even support an affirmative obligation on the government to undertake reforms and regulations to ensure the viability of a news ecosystem. This notion of a positive First Amendment, developed repeatedly by scholars and commissions, appears in the reasoning and results of some judicial decisions and deserves recognition and action in light of the demands of democracy under serious stress.”)

frameworks in related domains. For instance, in the field of neuro-technology, the uncertainties and risks associated with the development and deployment of drugs and devices have been mitigated through FDA reviews and the professional ethics of doctors and researchers. Additionally, aforementioned IRBs provide oversight by reviewing and monitoring research protocols involving human subjects to minimize psychological, physical, and material harms to subjects.

Drawing from the models of FDA approval, a framework for prior risk assessment, informed consent, the minimization of harm principles, and continuous review of safety could be implemented in the context of AI systems. Individual AI labs could establish independent ethics departments governed by industry-wide rules, akin to the role of IRBs in academic research. This would ensure that the development and deployment of AI systems are subject to rigorous ethical scrutiny and ongoing monitoring to identify and mitigate potential risks to users' autonomy and well-being.

The AI industry could benefit from the establishment of professional ethics for AI engineers and self-regulatory mechanisms. In *AI's Hippocratic Oath*, Professor Chinmayi Sharma argue that professionalization of AI engineering, academic requirements, licensing, codes of conduct, disciplinary actions, and malpractice liability, would require AI engineers to consider ethical implications and potential societal harms before building AI systems.¹⁵⁷ Sharma finds this approach might overcome roadblocks to traditional regulation by conscripting the technical experts themselves to set evolving standards rather than relying on less knowledgeable policymakers.¹⁵⁸

Industry-wide efforts, such as the Global Internet Forum to Counter Terrorism (GIFCT) for combating terrorist content or the PhotoDNA initiative for identifying child sexual abuse material (CSAM), have aimed to address harmful content online. Similar collaborative approaches could be adopted to tackle the challenges posed by AI-mediated information operations and manipulation. For instance, OpenAI's decision to ban certain state actors from using its platform demonstrates a proactive stance in preventing the misuse of AI for malicious purposes.¹⁵⁹ By establishing industry-wide standards, sharing best practices, and coordinating efforts to detect and mitigate AI-mediated harms, the AI community can work together to safeguard free speech values and protect users' autonomy.

¹⁵⁷ Chinmayi Sharma, *AI's Hippocratic Oath* 36-39 (Wash. U. L. Rev., forthcoming), <https://papers.ssrn.com/abstract=4759742>

¹⁵⁸ *Id.* at 46-48.

¹⁵⁹ OpenAI, *Disrupting Malicious Uses of AI by State-Affiliated Threat Actors*, OpenAI (Feb. 14, 2024), <https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors>

Furthermore, the concept of information fiduciaries and the duty of loyalty, which have been proposed in response to the exploitative power of social media platforms and search engines,¹⁶⁰ are particularly well-suited to the relationship between AI agents and users. There is a clear principal-agent relationship characterized by information asymmetry, where aligning the agent's behavior with the principal's interests is crucial. While social media platforms serve users like retail customers, AI agents cater to the personal preferences of users, resembling therapist-patient or attorney-client relationships. Adopting this framework could guide the reform of contractual relationships between AI companies and users, ensuring that AI systems are designed to prioritize users' interests and protect their autonomy.

In addition to facilitating knowledge-producing institutions, public investments may be necessary to develop small-scale AI systems that can diversify information sources and promote a more pluralistic information ecosystem. Moreover, investing in AI literacy programs in schools and non-profit organizations can empower individuals to critically evaluate the information they encounter and detect falsehoods or manipulative content generated by AI systems. Efforts should also be made to provide equitable access to generative AI tools and training, especially in underfunded public schools, to prevent disparities that could further disadvantage underserved communities.

Realizing free speech values will require a multi-pronged approach spanning structural risk assessment, professionals' self-regulation, public education, and a reconstruction of power structures. While not an exhaustive list, these measures collectively can help bridge the "free speech values gap" by complementing constitutional free speech protections with a proactive, systemic reinforcement of the core values underpinning free expression.

CONCLUSION

The freedom to think for oneself, to form beliefs and opinions based on a diverse range of information and perspectives, is essential not only for individual flourishing but for the functioning of a free and democratic society. However, advancements in generative AI systems and neurotechnology are

¹⁶⁰ Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 UC DL REV. 1183 (2015); Jack M. Balkin, *The Fiduciary Model of Privacy*, 134 HARV. L. REV. F. 11 (2020); Claudia E. Haupt, *Platforms as Trustees: Information Fiduciaries and the Value of Analogy*, 134 HARV. L. REV. F. 34 (2020); Neil Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 WASH. U.L. REV. 961 (2021); Woodrow Hartzog & Neil Richards, *Legislating Data Loyalty*, 97 NOTRE DAME L. REV. REFLECTION 356 (2022); and Woodrow Hartzog & Neil Richards, *The Surprising Virtues of Data Loyalty*, 71 EMORY LJ 985 (2021).

demonstrating novel capabilities to access, interpret, and even influence our innermost thoughts and cognitive processes. The sanctity of the human mind is no longer inviolable, but susceptible to being read and influenced by external forces.

Safeguard individual autonomy in the face of these insidious challenges requires proactive interpretation of fundamental rights. A multi-faceted approach is necessary. Privacy law must evolve beyond the traditional focus on secrecy and access control to encompass the regulation of how personal data is used to influence individuals' choices and behaviors. Concepts like intellectual privacy, data loyalty, and the right to informational self-determination provide valuable frameworks for addressing the unique risks posed by mind-reading and mind-manipulating technologies.

While the First Amendment's protections for freedom of thought and expression are not directly applicable to the private development and deployment of these technologies, they can still provide valuable guidance for regulating manipulative practices and preserving the conditions necessary for individuals to engage in free and independent thinking. I advocate envisioning broader social institutions including professional ethics, literacy education, and public funding towards proactively embodying free speech values through technology design, norms, and regulations.