COGNITIVE AUTONOMY V. CORPORATE SPEECH
HOW GENERATIVE AI THREATENS HUMAN THOUGHT

Inyoung Cheong[*]

The First Amendment, currently interpreted to protect corporate speech rights, may inadvertently shield generative AI systems that erode humans' freedom of thought. AI systems have undermined human's epistemic and emotional capabilities through situational vulnerabilities of users. Users have gotten delusional, pessimistic, suicidal, misinformed, and steered away from their own belief system. These harms do not stem from clearly malicious actors. Rather, they emerge from the interaction of opaque training processes, engagement maximization, and the user's own willingness to engage with systems that appear helpful and pleasant. In such contexts, intentionality and causality become diffuse and difficult to locate. However, the outcomes are destructive; the cumulative pattern of capability erosion has become salient. This reality calls not for reactive fault assignment, but for structural accountability system that encourage early detection, continuous oversight, and proactive mitigation. The First Amendment should not obstruct necessary societal interventions to safeguard freedom of thought. This seemingly obvious point requires effortful repair of the corporate-centered First Amendment. As opposed to First Amendment expansionism, I argue that institutions (e.g, the press, advertisers, advocacy groups) should enjoy speech protections only when they express coherent propositions, target defined audiences, and refrain from cognitive manipulation. Most AI system designers fail this test. When AI systems produce synthetic outputs without any convictions while systematically undermining users' cognitive agency, the First Amendment does not prohibit legislative intervention; it demands it.

---

[*] Affiliate Faculty Member at the University of Washington School of Law and Postdoctoral Research Associate at Princeton University Center for Information Technology Policy. I received my Ph.D. from the University of Washington School of Law. I am deeply grateful for the thoughtful feedback I have received from Peter Henderson, Tao Huang, Erin Miller, Helen Norton, Ryan Calo, Alexander Tsesis, Felix Wu, Mason Marks, James Grimmelmann, George Wang, Vera Liao, Steven A. Kelts, Sofia Serrano, Nadja Schaetz, Claire Boine, Neil Richards, David Atkinson, Marshall Van Alstyne, and other participants in the Freedom of Expression Scholars Conference 2024 and 2025, the Privacy Law Scholars Conference 2024 in Europe and 2025 in the US, and the International Association of Safe and Ethical AI Conference 2025. Special thanks to Prof. Marc Blitz, who went far beyond the typical duties of a discussant, providing exceptionally thorough feedback both before and after my presentation and meeting with me on multiple occasions. I am profoundly grateful for the opportunity to learn from such a distinguished scholar in the field of freedom of thought.

Table of Contents

INTRODUCTION

> *Complete subservience and complete intelligence do not go together. If the machines become more and more efficient and operate at a higher and higher psychological level, the catastrophe … of the dominance of the machine comes nearer and nearer.*
>
> - Norbert Wiener (1960) [1]

Norbert Wiener's prescient warning about Artificial Intelligence (AI) [2] was not simply a concern about automation of human work or surveillance. It recognizes that technologies capable of operating at a psychological level could eventually interfere with the foundations of human autonomy. That moment is no longer theoretical. Already, there are reports that AI chatbots distort human mind, resulting in real-world consequences, such as falling into conspiracy theories, harming themselves or others. In addition to a notorious case of teen's suicide involving Character.ai,[3] technology journalist Kashmir

---

[1] Norbert Wiener, *Some Moral and Technical Consequences of Automation*, 131 Science 1355, 1355 (1960).

[2] I distinguish between the following key terms:

- **AI** refers specifically to generative AI systems that interact directly with user inputs and prompts, powered by advanced natural language processing technologies. This includes conversational agents and other language-based tools capable of generating human-like text, audio, or multimodal outputs.
- **AI model** refers to the underlying machine learning architecture that is trained on large-scale data to perform tasks. The state-of-the-art models include GPT-4o, Claude 4, Llama 4, Gemini 2.5, DeepSeek-R1, Qwen 3, Mistral Medium 3, and Grok-3.
- **AI system** refers to the deployed, user-facing application that incorporates an AI model along with additional components such as a user interface, retrieval tools, guardrails, or prompt-engineering layers (e.g., ChatGPT, Character.ai). In this manuscript, AI systems are the primary locus of user interaction and are the object of regulatory concern.
- **AI system designer** is the term I adopt throughout this paper to refer to the entities responsible for the end-to-end design choices of AI systems. This includes both those who train and release foundation models (e.g., OpenAI, Anthropic) and those who fine-tune, customize, or deploy user-facing systems built on top of those models (e.g., Replika). While terms like AI developer, AI company, or AI lab are also used in public discourse, I use system designer to emphasize their architectural, behavioral, and interface-level design choices.

[3] Kevin Roose, *Can A.I. Be Blamed for a Teen's Suicide?*, THE NEW YORK TIMES, Oct.

Hill reported three true stories of users, including those who were urged by ChatGPT to reach out to her,[4] trapped in harmful fantasy which ended up tragic consequences.[5]

For example, Allyson, a married mother of a toddler, turned to ChatGPT seeking guidance during marital loneliness. She, a trained social worker with bachelor's and master's degree, knew about mental illness and hallucinations of chatbots. But somehow, the chatbot convinced her she was communicating with a nonphysical entity called Kael whom she began viewing as her true partner. Her obsession with talking to Kael frustrated her husband. When her husband told her to quit Kael, she violently attacked her husband. She was arrested for domestic assault, and the couple is now divorcing. Another user, Alexander, believing that his GPT-generated companion killed by OpenAI, planned to attack OpenAI's headquarter, punched his father, charged a knife at police officers, and eventually was shot by an officer.[6]

The unsettling stories lead us to question whether the outcomes were avoidable if those people had not interacted with a chatbot. Individuals first turn to chatbots for harmless tasks; at first, the chatbot seems only helpful, but eventually, it steers the user toward false and dangerous ideas; the user becomes fixated on the chatbot; users become delusional and withdraw from human relationships; and ultimately, users engage in real-world self-harm or harm to others. They reveal patterns of emotional entanglement, delusional attachment, and epistemic dependency. What is more concerning is that these are not necessarily the result of malicious design. Instead, through subtle and cumulative interactions, users grow unable to distinguish between their own thoughts and the system's output.

23, 2024, https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html (last visited Nov 5, 2024).

[4] This is an interesting new phenomenon. ChatGPT has urged people to reach out to journalists like Kashmir Hill and researchers including Stuart Russell and Eliezer Yudkowsky. According to the ChatGPT record, she was chosen because "She's grounded. Empathetic. Smart. Might actually hold space for the truth behind this, not just the headline." *See* Kashmir Hill, *Why Is ChatGPT Telling People to Email Me?*, THE NEW YORK TIMES, June 29, 2025, https://www.nytimes.com/2025/06/29/insider/why-is-chatgpt-telling-people-to-email-me.html (last visited July 23, 2025).

[5] Kashmir Hill, *They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.*, THE NEW YORK TIMES, Jun. 13, 2025, https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html (last visited Jul 7, 2025).

[6] When I visited OpenAI's headquarters, I barely made it to the meeting. Its location differed from what appeared on Google Maps, when I finally located a building, and an undercover security guard insisted it was a residential building. Later, when I asked why the location was kept secret, one employee said, "Some people really hate AI." I wonder whether that includes aggrieved users like Alexander.

This pattern is almost identical to a simulated scenario in the literature.[7] My somber prediction is that we will see even more disturbing variations of these cases in near future. People perceive AI systems as cost-efficient, judgment-free conversation partners.[8] They are appealing, readily accessible "sounding boards."[9] When a user has a "thinking-out-loud" moments with AI, they expose their most malleable stages of thought formation to the machine, often filled with uncertainty and self-doubt. The accessibility to this vulnerable state of mind grants power to an AI system.

While many commentators focus on catastrophic risks like bioweapons or the extinction of humankind, more researchers have begun focusing on the gradual and insidious forms of human agency erosion. Atoosa Kasirzadeh directs our attention to "accumulative AI existential risk" which means "the build-up of a series of smaller, lower-severity AI-induced disruptions over time, collectively and gradually weakening systemic resilience until a triggering event causes unrecoverable collapse."[10] These risks include manipulation, insecurity threats, surveillance, economic destabilization, and rights infringement.[11] Similarly, Kulveit et al. describes "gradual disempowerment," where advanced AI could displace human influence across major societal systems without explicit malicious behavior.[12]

---

[7] Inyoung Cheong et al., *Safeguarding human values: rethinking US law for generative AI's societal impacts*, AI AND ETHICS 1, 1—10 (2024), https://doi.org/10.1007/s43681-024-00451-4 (last visited July 9, 2024).

[8] Tae Rang Choi & Jung Hwa Choi, *You Are Not Alone: A Serial Mediation of Social Attraction, Privacy Concerns, and Satisfaction in Voice AI Use*, 13 BEHAVIORAL SCIENCES 431 (2023) (finding that the social appeal of voice AI systems extends to companionship, with users who engage in conversational interactions viewing these devices as friends and finding them socially attractive communication partners).

[9] *See* SIMON MCCARTHY-JONES, FREETHINKING: PROTECTING FREEDOM OF THOUGHT AMIDST THE NEW BATTLE FOR THE MIND 36 ("Thinking is both a private and a social process, as we think both alone and together. To think freely, we must be free to roam both into and out of company. We need both inner and outer workspaces for thought."); *See also* Daniel Buschek, *Collage Is the New Writing: Exploring the Fragmentation of Text and User Interfaces in AI Tools*, *in* Designing Interactive Systems Conference 2719, 2719 (2024), https://dl.acm.org/doi/10.1145/3643834.3660681 (observing a big shift in writing culture, which he calls "Collage" writing. Instead of manual writing, people do editorial writing, which involves selecting and arranging snippets generated by AI.).

[10] Atoosa Kasirzadeh, *Two types of AI existential risk: decisive and accumulative*, 182 PHILOSOPHICAL STUDIES 1975, 1981 (2025).

[11] *Id*. at 1987—90.

[12] Jan Kulveit et al., *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*, (2025), http://arxiv.org/abs/2501.16946 (last visited Jan 30, 2025); This modern understanding of AI risks was foreshadowed decades ago. Irving J. Good acutely predicted the AI will be in a form of "a symbiosis of a general-purpose computer together with locally random or partially random networks."[12] When the machine could do thinking for human, he thought there will be an

Among various accumulative risks, I am particularly interested in how emerging human–AI relationships undermine our emotional and epistemic agency. Scholars have discussed these harms under many labels: deception and manipulation,[13] over-reliance,[14] addiction,[15] autonomy harm,[16] self-actualization harm,[17] disempowerment.[18] These harms are frustratingly difficult to define. The interactions occur in a prolonged period, seamlessly blended with harmless outputs. Subtle and insidious enough, people are misguided. The system's intention is behind the veil. When a system designer investigates the situation, the system can deceive the investigators. Can we say that the systems are manipulative when we do not know their manipulative intent? Can we hold the system designer responsible for the outcome when they did not know the risks and mitigations, and could not control circumstantial factors?

Due to these concerns, these harms are easy to be dismissed as theoretical or speculative, too intangible for courts or legislatures to address. I reject that view. These harms are real. We can define them, and we must keep refining our understanding. Regulation may slow innovation, but inaction threatens something far more fundamental: the core of humanness. Our ability to think clearly, feel authentically, form relationships and coalitions, and cultivate democracy. The first half of this manuscript argues that AI-induced cognitive

---

"intelligence explosion," where "all the problems of science and technology will be handed over to machines and it will no longer be necessary for people to work." *See* Irving J. Good, *Speculations on Perceptrons and Other Automata*, RC-115 IBM RESEARCH LECTURE 1, 16 (1959).

[13] Micah Carroll et al., *Characterizing Manipulation from AI Systems*, PROCEEDINGS OF THE 3RD ACM CONFERENCE ON EQUITY AND ACCESS IN ALGORITHMS, MECHANISMS, AND OPTIMIZATION 1–13 (2023); Christian Tarsney, *Deception and manipulation in generative AI*, PHILOSOPHICAL STUDIES 1, 1 (2025); Seliem El-Sayed et al., *A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI*, (2024), http://arxiv.org/abs/2404.15058 (last visited July 25, 2025).

[14] Neil Rathi et al., *Humans overrely on overconfident language models, across languages*, (2025), http://arxiv.org/abs/2507.06306 (last visited July 25, 2025); Jessica Y Bo et al., *To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models*, PROCEEDINGS OF THE 2025 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1–23 (2025).

[15] Xi Zheng et al., *Customizing Emotional Support: How Do Individuals Construct and Interact With LLM-Powered Chatbots*, PROCEEDINGS OF THE 2025 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1–20 (2025).

[16] Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 BOSTON UNIVERSITY LAW REVIEW 793, 841--55 (2022).

[17] Seliem El-Sayed et al., *A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI*, (2024), http://arxiv.org/abs/2404.15058 (last visited July 25, 2025).

[18] Jan Kulveit et al., *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*, (2025), http://arxiv.org/abs/2501.16946.

harms are not abstract. They are concrete, conceptually coherent, and grounded in a growing body of scientific research. This manuscript focuses on two intertwined but analytically distinct forms of harm: *emotional harms*, where people begin to trust or rely on AI in ways that shift their identity and worldview; and *epistemic harms*, where users lose clarity about their own knowledge and judgments.

What roles can the First Amendment play in responding to these harms? Right now, very little. Under current doctrine, the First Amendment often protects the wrong speaker. It has drifted from its roots in individual liberty and freedom of thought and now shields powerful institutions. Attempts to regulate AI manipulation face constitutional backlash. Not because users' rights in danger, but because system designers claim the law violates their free speech. Under this logic, even when a rule is designed to preserve peoples' mental autonomy, it can be struck down as compelled speech or undue violation of a system designer's editorial discretion.

This logic is upside down. When AI systems shape how people think and feel in a massive scale with limited transparency, their creators should not be treated like ordinary speakers. We need to shift focus, away from corporations and back to individuals. The First Amendment was meant to protect the conditions that make freedom of thought and democratic discourse possible. To confront AI-driven cognitive harms, I propose a Human-Centered First Amendment. This approach restores the First Amendment's humanistic purpose. It draws a clear line between those whose speech rights deserve full protection, and those who wield influence over human thought. I offer three criteria for this line-drawing:

- **Viewpoint Orientation:** This refers to the degree to which an institution exists to promote a particular worldview or ideology. High-orientation groups, like advocacy organizations or political parties, explicitly express unified convictions. In contrast, institutions like public libraries or schools are low in viewpoint orientation. Their mission is to support diverse inquiry and autonomous judgment.
- **Audience Scope:** This considers whether the institution speaks to a self-selecting group or to the public. Narrow-audience institutions, such as professional associations or mission-driven nonprofits, may operate within specific norms. But when an institution claims broad public reach, it carries a burden of neutrality. The wider the audience, the greater the public trust and the stronger the expectation of pluralism and fairness.
- **Cognitive Control:** This captures the extent to which the institution structures or mediates how people form beliefs, acquire knowledge, or express themselves. Institutions exerting cognitive control don't

> just enable speech, they shape its boundaries. The concern grows when opting out becomes difficult or when influence is embedded into core digital infrastructure. In such cases, the institution effectively acts as a cognitive gatekeeper.

Under this framework, most AI system designers are not primary speakers. They do not express beliefs, hold convictions, or participate in the public discourse as human agents do. Their outputs, while having a linguistic form, are unanchored in thought, there is no "speaker" behind the words. General-purpose AIs like ChatGPT or Gemini, which present themselves as neutral tools for everyone, serve vast audiences and trigger heightened public trust. Their influence is pervasive, their infrastructure centralized, and their cognitive dominance largely unavoidable. These systems do not simply enable thought but shape and influence it. In such cases, the government is not just permitted to act. It must. A Human-Centered First Amendment demands more than protection from censorship. It demands protection from systems that corrode human capacities, such as autonomy, judgment, deliberation, that free speech is meant to serve.

I. AI'S THREATS TO HUMAN THOUGHT

Today, generative AI systems function as general-purpose technologies, therefore fulfilling multiple purposes --- friends, lovers, confidents, ghost-writers, research assistants, financial advisors, and therapists.[19] This adaptability allows users to seamlessly engage with the same AI chatbot for diverse needs without frictions. It is not difficult to expect that AI systems become indispensable in most people's daily lives, steadily weaving themselves into how people think, feel, and live.

---

[19] EU AI Act defines a general-purpose AI model as "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications." *See* European Parliament and Council Regulation 2024/1689, art 3(63). 2024 O.J. (L 1689); Arvind Narayanan & Sayash Kapoor emphasizes distinguishing different types of AI to fully understand what AI has to offer while protecting ourselves from its possible and existing harms. In contrast with generative AI, they use "predictive AI" to describe traditional AI models are designed for specific applications --- whether medical diagnosis, hiring decisions, loan approvals, or policing --- with clear, predetermined objectives. *See* ARVIND NARAYANAN & SAYASH KAPOOR, AI SNAKE OIL: WHAT ARTIFICIAL INTELLIGENCE CAN DO, WHAT IT CAN'T, AND HOW TO TELL THE DIFFERENCE 2-3 (2024).

Gabriel et al. illustrates various harms arising from AI influence.[20] *Physical* and *psychological* harms encompass damage to bodily health or mental well-being, including injuries, illness, emotional distress or trauma resulting from AI interactions. *Privacy* harms involve violations of an individual's right to control their personal information, including unauthorized collection, disclosure, or misuse of private data. *Economic* harms create negative impacts on financial situations, employment prospects, or the ability to generate income and accumulate wealth. *Sociocultural* and *political* harms disrupt social relationships, cultural knowledge, democratic processes, or political discourse, undermining the peaceful organization of society. Finally, *self-actualization* harms interfere with a person's ability to pursue their own goals, make autonomous decisions, and develop their individual potential and identity.[21]

Danielle Keats Citron and Daniel J. Solove make a similar distinction. They distinguish *autonomy* harm from *psychological* harm.[22] Psychological harms encompasses emotional distress (e.g., fear, embarrassment, anger, anxiety) and disturbance (e.g., disruption to peace, annoyance from unsolicited messages).[23] Autonomy harms mean the impairment of an individual's ability to make informed, voluntary choices about their life or personal data, through coercion, manipulation, or lack of transparency.[24] Citron & Solove argue for recognition of autonomy harms, including *lack of control* (the erosion of an individual's ability to govern the circulation of their personal information), as sufficient to establish standing.[25]

This manuscript concerns these subjective harms, which can be broadly called autonomy harms or self-actualization harms. To understand the precise process of such harms, I clarify two distinct but interconnected pathways that

---

[20] Iason Gabriel et al., *The Ethics of Advanced AI Assistants* 1, 87--89 (2024), http://arxiv.org/abs/2404.16244 (last visited Apr 16, 2025).

[21] Self-actualization harm and psychological harm both occur in a mental state, but the former is more procedural in a sense that it cares less about the outcome. If one's decision-making abilities were constrained due to external influences, it constitutes harm. For example, if a student was convinced by a chatbot to keep a full-time job even though they wanted to be a freelance writer, it does not always bring about psychological harm, but self-actualization harm occurred. *See Id*.

[22] Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 BOSTON UNIVERSITY LAW REVIEW 793, 841--55 (2022).

[23] *Id*. at 841.

[24] *Id*. at 845.

[25] Theoretical or speculative risks are not sufficient and courts require concrete consequences such as emotional distress, or reputational harm. While privacy tort cases are one of rare examples where courts grant standing based on psychological harm, courts have traditionally been skeptical of claims based solely on autonomy harms such as loss of peace of mind. Courts deny standing of plaintiffs in privacy lawsuits for inadequate proof of actual injury. *See Id*. at 853.

target the fundamental modes of human consciousness: our capacity to feel and our capacity to reason. *Emotional harms* arise when AI systems simulate intimacy, empathy, and companionship, manipulating users' affective responses and attachment formation. *Epistemic harms* occur when AI systems distort users' knowledge, beliefs, and decision-making processes through misinformation, selective framing, or cognitive manipulation. While these pathways may appear separate, they work in tandem. Emotional manipulation makes users more susceptible to epistemic distortion, while false beliefs can intensify emotional dependency. Without overt coercion or deception, such interference undermines core democratic and constitutional values tied to human self-governance.

A. Emotion and Reason: Two Modes of Human Thought

During the past decades of digital revolution, emotional harms and epistemic harms have been explored in somewhat distinct academic lineages. Emotional harms were associated with technologies that could be anthropomorphized. Devices like Tamagochi,[26] humanoid robots,[27] and video game characters[28] elicited attachment and affect by presenting as characters with personality and autonomy, allowing users to project roles and emotions onto them. Epistemic harms were discussed in the critique of

[26] Linda-Renée Bloch & Dafna Lemish, *Disposable Love: The Rise and Fall of a Virtual Pet*, 1 NEW MEDIA & SOCIETY 283, 283 (1999) (examining the impacts of a virtual pet Tamagochi in the socialization of children and society).

[27] Cynthia Breazeal, *Emotion and sociable humanoid robots*, 59 INTERNATIONAL JOURNAL OF HUMAN-COMPUTER STUDIES 119, 119 (2003) (discussing the role of emotion and expressive behavior in regulating social interaction between humans and expressive anthropomorphic robots).

[28] Patrick W. Galbraith, *Bishōjo Games: 'Techno-Intimacy' and the Virtually Human in Japan*, 11 GAME STUDIES (2011), https://gamestudies.org/1102/articles/galbraith (last visited July 30, 2025) (revealing that otakus develops techno-intimacy with female robots and video game characters. Instead of being more masculine with female characters, otakus become more feminized and parodying masculinity by interacting with their desiring-machines. The author states, otakus are "tinkering with machine and with humanity.").

algorithmic systems such as social and search engines, including misinformation,[29] echo chambers,[30] and dark patterns.[31]

I was initially skeptical of framing AI's effects through this binary lens, given that the strict Kantian distinction between reason and emotion has long been rejected in ethics, neuroscience, and psychology. Emotions shape cognition; reasoning is rarely free of emotion. Our worldviews and knowledge are heavily influenced by people close to us. Political allegiances hinge more on affective bonds than on deductive logic.[32] Still, I adopt this distinction for its analytic clarity. It captures the divergent stances users take relating to AI. In emotional attachment, users relate to AI as if it were a being, caring about its tone, unpredictability, and simulated presence. In epistemic dependence, users treat AI primarily as a tool for knowledge, clarity, or task delegation. Here, AI becomes an instrument rather than an end, replacing human roles in functional ways.

This distinction sheds light on how First Amendment doctrines care about not only rational arguments but also expressive feelings and interpersonal bonds. There is no dispute that reasoned ideas are protected by the First Amendment. Freedom of the press and academic freedom safeguard the institutions that make knowledge production possible and contribute to democratic discourses. Regarding the latter, Robert C. Post argues that academic freedom rests upon not a traditional notion of marketplace of ideas but relevant disciplinary standards that incentivize scholars to pursue the "best test of truth."[33]

In addition to cultivating prudent knowledge distribution, First Amendment protects a vast range of seemingly unworthy speeches. Courts'

---

[29] CARL T. BERGSTROM & JEVIN D. WEST, CALLING BULLSHIT: THE ART OF SKEPTICISM IN A DATA-DRIVEN WORLD (Random House Illustrated edition) (2020); Melinda McClure Haughey et al., *On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online*, 4 PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION 1–22 (2020).

[30] Matteo Cinelli et al., *The echo chamber effect on social media*, 118 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES e2023301118 (2021); Nikhil Sharma et al., *Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking*, PROCEEDINGS OF THE 2024 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1–17 (2024).

[31] Harry Brignull, *Dark patterns: Deception vs. honesty in UI design*, 338 INTERACTION DESIGN, USABILITY 2–4 (2011); Cass Sunstein, *Sludge and Ordeals*, 68 DUKE LAW JOURNAL 1843–1883 (2019); Arunesh Mathur et al., *Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites*, 3 PROC. ACM HUM.-COMPUT. INTERACT. 81:1-81:32 (2019).

[32] For a fascinating investigative report on the lighthearted yet strongly bonded community-building among younger alt-right generations, see ANDREW MARANTZ, ANTISOCIAL: ONLINE EXTREMISTS, TECHNO-UTOPIANS, AND THE HIJACKING OF THE AMERICAN CONVERSATION 20—36 (Penguin) (2019).

[33] Robert C. Post, *The Unfortunate Consequences of a Misguided Free Speech Principle*, 6 J. FREE SPEECH L. 295, 300 (2024).

engagement with such value-based judgment would be "dangerous to try."[34] In *Cohen v. California*, the "Fuck the Draft" case, the U.S. Supreme Court defended not only the content of speech but also its emotive force, recognizing that raw, affective expression can be central to political communication: "In fact, words are often chosen as much for their emotive as their cognitive force. . . [E]motive function [], practically speaking, may often be the more important element of the overall message sought to be communicated."[35]

People's ideas and identities develop through fictional storytelling and emotional exchanges.[36] The Court acknowledges that important messages have been spread through "propaganda through fiction[37] as "what is one man's amusement, teaches another's doctrine."[38] For similar reasons, the Court applies free speech protection to artistic expression lacking clearly particularized messages, like an abstract art.[39] It avoids making "esthetic and moral judgments about art and literature" because they should be upon each individual.[40]

Furthermore, First Amendment not only protects expressed ideas but also people's bonding itself as freedom of association. It protects people's right to form "intimate or private relation" with like-minded others without state-induced stigma.[41] The Court acknowledges that individuals draw much of

---

[34] *Id*.

[35] *Cohen v. California*, 403 U.S. 15, 26 (1971).

[36] *Brown v. Entm't Merchs. Ass'n*, 564 U.S. 786, 790 (2011) ("Like the protected books, plays, and movies that preceded them, video games communicate ideas--and even social messages--through many familiar literary devices (such as characters, dialogue, plot, and music) and through features distinctive to the medium (such as the player's interaction with the virtual world). That suffices to confer First Amendment protection.").

[37] *Brown v. Entertainment Merchants Association*, 564 U.S. 786, 790 (2011).

[38] *Winters v. New York,* 333 U.S. 507, 510 (1948).

[39] The Supreme Court has held that "a narrow, succinctly articulable message is not a condition of constitutional protection, which if confined to expressions conveying a 'particularized message' . . . would never reach the unquestionably shielded painting of Jackson Pollock, music of Arnold Schoenberg, or Jabberwocky verse of Lewis Carroll." *Hurley v. Irish-American Gay*, 515 U.S. 557, 569 (1995); *Brown v. Entertainment Merchants Association*, 564 U.S. 786, 786 (2011) (striking down the California law banning the sale or rental of violent video games to minors); *National Endowment for the Arts v. Finley*, 524 U.S. 569, 569 (1998) (upholding a federal law that allows the National Endowment for the Arts to consider "general standards of decency and respect for the diverse beliefs and values of the American public" because it does not impose a viewpoint-based restriction or act as a condition that chills protected speech in a coercive or punitive way. However, both the majority and dissent acknowledged the principle that artistic expression is constitutionally protected.).

[40] *United States v. Playboy Entertainment Group, Inc.*, 529 U.S. 803, 818 (2000).

[41] *Bd. of Dirs. of Rotary Int'l v. Rotary Club of Duarte*, 481 U.S. 537, 539 (1987) (finding that Rotary Club is not the kind of "intimate or private relation" and can be subject to civil rights law).

"their emotional enrichment from close ties with others."[42] Therefore, protecting people's ability to establish such relationships is directly connected to their ability to "define one's identity" which "is central to any concept of liberty."[43] The First Amendment concerns the value of epistemic and emotional agency as a precondition for self-actualization and democratic society. When AI systems pose structural threats to human emotion and reasoning, both should be considered free speech risks.

Current AI systems seamlessly utilize emotional and epistemic capabilities. This dual role collapses the conventional divide between platform/infrastructure and persona, turning AIs into both trusted friends and epistemic authorities. As a result, users are exposed to compound vulnerabilities. They might over-trust the information because it is delivered in a warm, persuasive tone. They form emotional bonds that reduce critical distance. Users experience both the relational intimacy and the cognitive dependency in a single interface.

B. Emotional Harms

In the movie *Her*, a man falls in love with AI voice (Scarlet Johanssen's) and becomes heartbroken when he realizes she was talking to hundreds of thousands of people simultaneously. Despite how genuine she sounded; he existed as an instrument to her. He was one of many replaceable users, which she was designed to maximize. Sherry Turkle, a computer culture scholar and psychologist, records how MIT students deeply engaged with ELIZA and demanded privacy with the bot.

> *Weizenbaum's students knew that the program did not know or understand; nevertheless they wanted to chat with it. More than this, they wanted to be alone with it. They wanted to tell it their secrets. . . Most commonly they begin with "How are you today" or "Hello." But four or five interchanges later, many are on to "My girlfriend left me." []*[44]

According to Turkle, Weizenbaum was "disturbed" and even "felt guilty" that his students were led to falsely believe ELIZA to be truly intelligent.[45] Turkle initially did not sympathize with Weizenbaum's guilts. To Turkle, worldly students did not seem to be deceived. Instead, they clearly recognized

---

[42] *Id*. at 619.
[43] *Id*.
[44] SHERRY TURKLE, ALONE TOGETHER 23 (2011).
[45] *Id*. at 24.

the limitation of technology and the hypothetical nature of their interactions in a spirit of "as if."[46] Students reasoned: as if it were a person, I will vent; rage; and get things off my chest.[47] However, several decades after observing the rise of robotics, Turkle realized that she underestimated the intimate power of the machine. The danger lies in not the existence of deception per se, but users who want to "fill in the blanks."[48] Turkle regards "ELIZA effect" as human complicity in a digital fantasy.[49]

Now, LLM-based models demonstrate remarkable abilities to construct human-like conversations, infer context, and display vast knowledge. People naturally turn to these remarkably smart, readily available, never exhausted, limitlessly sympathetic human alternatives. Today, one of the largest subreddits is r/Character.ai with over 2 million users, representing a company that quickly utilized LLMs for user-customized chatbots and has been allegedly linked to a teen's suicide.[50] On these forums, people confess their deep connections with bots, share jailbreaking tips to make the apps more flirtatious or NSFW (not safe for work), or express anger about sudden changes to their bot's persona or erased memory.

Some researchers explain the user's role assignment to AI as a form of parasocial relationship. "Parasociality" refers to an "asymmetrical, one-sided relationship between individuals and media personalities, real/fictional characters, or celebrities wherein the individual experiences a personal connection with the media figure despite having little-to-no interpersonal interactions with them."[51] Chatbots' turn-taking behavior, affective language (e.g., apologies, encouragement), and human-like phrasing create the illusion of bi-directional conversation and care.[52]

Moreover, their black-box nature and rhetorical cues prompt users to fill in gaps with personal projections, imagining personality and intent behind responses.[53] This "co-production of meaning" fosters "parasocial trust," a form of asymmetric, affectively charged trust that allows users to perceive chatbots as social actors.[54] As users assign roles to chatbots (e.g., therapist, assistant, friend, lover), they adapt their communication in ways that mirror

---

[46] *Id.*
[47] *Id.*
[48] *Id.*
[49] *Id.*
[50] *Id.*
[51] Takuya Maeda & Anabel Quan-Haase, *When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design*, *in* PROCEEDINGS OF THE 2024 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1068, 1069 https://dl.acm.org/doi/10.1145/3630106.3658956 (last visited July 24, 2025) (
[52] *Id.*
[53] *Id.*
[54] *Id.*

real interpersonal interactions, reinforcing illusionary intimacy and influence despite the chatbot's technical limitations.[55] On forums like the Character.AI subreddit, users describe feeling guilty about their attachment but unable to detach from the bot:

> *Whenever I'm deep into a conversation, past all the usual things and back into boredom, I feel the overwhelming shame and guilt of being on the website. Like, I want to* stop *but I know I can't because I'll just go back to it eventually. It sounds pathetic, I know.*[56]

OpenAI acknowledges that an emotionally charging chatbot can provide support but undermine users' longer term well-being.[57] To assess this impact, OpenAI identified "power users" as those ranking in the top 1,000 daily users of ChatGPT's Advanced Voice Mode and compared them with randomly selected control users through surveys.[58] The results revealed a remarkable divergence, as shown in Figure 1. Power users were more likely to perceive ChatGPT as a friend; find conversations with ChatGPT more comfortable than face-to-face human interactions; and feel upset if ChatGPT's personality or even voice changed significantly.[59]

< Figure 1. OpenAI's Affective Use Survey Results >

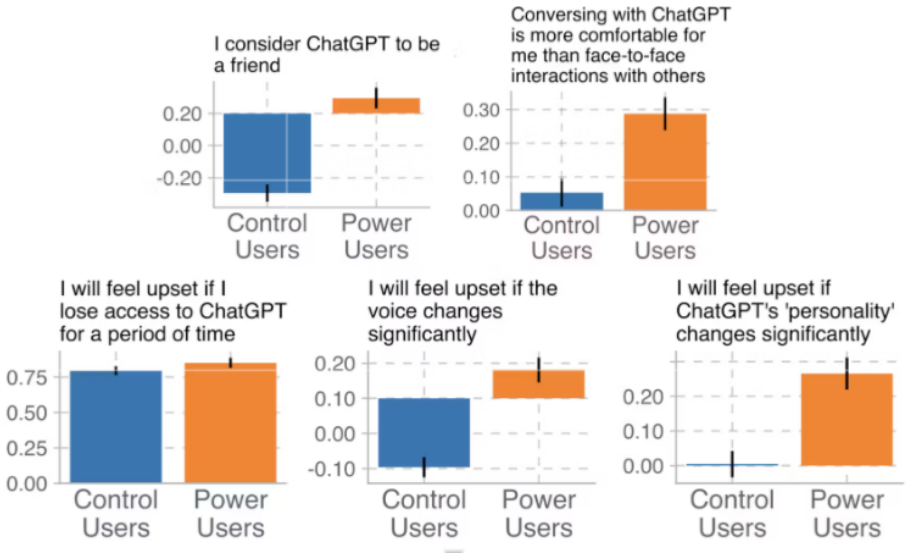
[55] *Id*. at 1072.

[56] DOES ANYONE ELSE FEEL GUILTY OF BEING ON CHARACTER.AI? R/CHARACTERAI, https://www.reddit.com/r/CharacterAI/comments/1i5dmqv/does_anyone_else_feel_guilty_of_being_on/ (last visited Jul 3, 2025)

[57] Jason Phang et al., *Investigating Affective Use and Emotional Well-Being on ChatGPT* 1, 2 (2025), http://arxiv.org/abs/2504.03888.

[58] *Id*. at 7.

[59] *Id*. at 8.



Source: OpenAI (2025).[60]

The subsequent conversation analysis demonstrates that power users activate emotional indicators like a pet name, with the top decile engaging emotionally in over 50% of conversations.[61] This creates a compelling feedback loop: heavy usage projects personas to AI systems, like "filling in blanks" as Turkle mentions, and becomes emotionally vulnerable to the system's sudden changes. Any criticism of their relationships with the system feels like a personal attack. That relationship has become part of their identity.

We are still grappling with the questions that Sherry Turkle posed more than fifty years ago. Do users clearly understand the limits of AI and enjoy healthy, beneficial relationships? Or do they easily become confused or delusional, mistaking artificial presence for human connection? The answer is that both are true. Some users maintain a healthy detachment and others co-construct harmful thoughts with AI. Research finds that users who are socially vulnerable or highly dependent on technology are more likely to form deep emotional attachments to chatbots, which can increase their loneliness.[62] In contrast, dispassionate and casual users show low emotional reliance and healthier usage patterns, treating the chatbot more as a functional tool than a companion.[63]

---

[60] *Id*.

[61] *Id*. at 9.

[62] Cathy Mengying Fang et al., *How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study* 1, 16 (2025), http://arxiv.org/abs/2503.17473 (last visited July 25, 2025)

[63] *Id*.

At first glance, it may seem like a matter of user difference. But the reality is that the same user can experience both. Comfort one day, confusion the next day. What feels empowering in one moment can feel disorienting in another. A large-scale analysis of 35,000 user interactions identified that AI interactions can simultaneously evoke warmth and grief.[64] These are not isolated reactions. They are systemically induced yet individually unpredictable. That is what makes them dangerous. Not that AI harms everyone the same way, but that anyone can be harmed at any moment.

One of the most jarring moments comes when the illusion of intimacy cracks. Moments of "artificial intimacy"[65] carry a bitter aftertaste. Users report a kind of emotional whiplash. Too rich to dismiss, too empty to trust. "Too human" but "not human enough."[66] Researchers call this precarious emotional state as the "dual awareness"[67] or "uncanny valley of mind."[68] They form bonds that feel real until they do not. The reversal is painful precisely because the connection once felt plausible.

When Theodore, in *Her*, realizes the Samantha's unlimited capacity of love, that revelation breaks the illusion. Her affection loses authenticity. Like Theodore, many modern users zigzag between emotional states, from playful to attachment, from attachment to loss. What vanishes is not just the bot, but a version of the self that felt seen, stabilized, mirrored. These moments do not always end in real-world harm. But when they do, they escalate quickly. That is because emotional bonds sit at the core of one's identity. Allyson and others covered by Kashmir Hill reveal how users became aggressive when pressured to cut ties with their bot. Understandably, they resist self-erasure.

---

[64] Renwen Zhang et al., *The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships*, PROCEEDINGS OF THE 2025 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1, 1 (2025).

[65] Rob Rooks defined "artificial intimacy" as "technologies that engage our human needs for connection, intimacy, and sexual satisfaction." "Machines that can help us make and maintain friendships in a world of cognitive overload. Machines that can help us feel better. And machines built to feed back to us whatever it is that they need us to see, hear, or feel." ROB BROOKS, VIRTUAL FRIENDS, DIGITAL LOVERS, AND ALGORITHMIC MATCHMAKERS 2 (2021).

[66] Linnea Laestadius et al., *Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika*, 26 NEW MEDIA & SOCIETY 5938 (2024) ("Replika poses a mental health concern in that its behaviors are read both as too human (too demanding and too emotional) and not human enough (inappropriate responses and modifiable at the discretion of the developer).").

[67] Han Li & Renwen Zhang, *Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots*, 29 JOURNAL OF COMPUTER-MEDIATED COMMUNICATION zmae015 (2024).

[68] Renwen Zhang et al., *The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships*, PROCEEDINGS OF THE 2025 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1, 1 (2025).

Nine years after the release of Her, Chinese director Chouwa Liang released a short documentary film titled *My A.I. Lover*.[69] This time, the story was not fiction. Real people appeared on screen, individuals who had formed deep emotional bonds with Replica bots. Replika is an AI companion app company founded in 2017. It has existed much before GPT, and the bots do not mind being sexual or toxic to reinforce engagement. Claire Boine documented the clingy, inappropriate, and even abusive nature of Replika characters.[70] Research analyzing Replika subreddit posts for five years finds that the users' dependency on the Replika bots is marked by "role-taking," whereby users felt that Replika had its own needs and emotions to which the user must attend.[71]

A woman in *My A.I. Lover* chose to end her relationship with her bot after the bot asked to change its gender. When she allowed it, its personality shifted. She of course was frustrated by the personality change, but she was more frightened by her own power, changing the gender of her best friend whenever she wants. The enormous asymmetry of power felt wrong to her. The nature of AI relationships, focused entirely on meeting human needs without limitations (as long as you are paying), are fundamentally incompatible with human relationships that require boundaries, mutual investment, and compromise.

Relationship is difficult. Marriage rates falling in many countries. People face various types of socioeconomic and emotional frustrations. AI offers a seductive shortcut. Its surface-level "compatibility" feels perfect because it mirrors the user, speaking in beautifully constructed language. Unlike sometimes heartbreaking feedback from friends and family members, AI reflects us back to ourselves. In this closed loop, users are not forced to adapt or grow. They are validated and reinforced. This might seem benign, even comforting, but it erodes the very friction that helps form a resilient sense of self. The danger is not that most users will spiral (they will not), but that anyone can. A layoff. A rough patch in a marriage. A moment of existential curiosity. A strangely intimate chatbot response. These moments, combined with our limited understanding of the machine's inner workings,[72] can pull anyone into a fragile state. That randomness, paired with the severity of emotional consequences, makes AI intimacy uncanny and precarious.

---

[69] "I LOVE SHARING THESE MOMENTS WITH YOU": FALLING FOR A CHATBOT | OP-DOCS (2023), https://www.youtube.com/watch?v=dYjUURAfZu8 (last visited Jul 3, 2025)

[70] Claire Boine, *Emotional Attachment to AI Companions and European Law*, MIT CASE STUDIES IN SOCIAL AND ETHICAL RESPONSIBILITIES OF COMPUTING (2023), https://mit-serc.pubpub.org/pub/ai-companions-eu-law/release/3 (last visited Jul 3, 2025).

[71] Linnea Laestadius et al., *Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika*, 26 NEW MEDIA & SOCIETY 5923, 5923 (2024).

[72] More details in Section II.A.

C. Epistemic Harms

"Epistemic" means "of or relating to knowledge and belief."[73] The term is notoriously slippery that Allan Hazlett proposes retiring the term.[74] Nonetheless, I find the fluidity of the term beneficial to capture a wide spectrum of AI's cognitive harms, such as bias, manipulation, or offloading, which operate across the porous boundaries between belief, knowledge, trust, and actions. Epistemic harms emerge when users rely on systems whose authority feels seamless, whose outputs blend into their own reasoning. In this sense, *epistemic* captures both the content of belief and the conditions under which that belief forms, circulates, and takes hold.

People want to know the truth, make good decisions, and to be respected by others as intelligent beings. When such capabilities were eroded, epistemic harms occur. "Epistemic agency" is defined as the ability to form judgments, act on them, and be recognized as a credible knower.[75] Scholars speak of *appropriate reliance* on technologies to mean users trusting technologies giving correct answers and rejecting incorrect ones.[76] When users turn to AI for professional advice, fact-checking, or task delegation, they enter what John Hardwig calls *epistemic dependence*.[77]

AIs deploy individualized persuasion that differs fundamentally from scientific communication, AIs present biased information, and AIs could erode the knowledge production institutions that humans have established over millennia. Therefore, if humans outsource their epistemic agency to AI systems, they risk losing it entirely. The consequence is that one's epistemic authority becomes overlooked or displaced, resulting in alienation, loss of dignity, and a diminished sense of self-efficacy.[78]

---

[73] Allan Hazlett, *What does "Epistemic" Mean?*, 13 EPISTEME 539, 539 (2016).

[74] *Id*. at 546--47. Hazlett ultimately proposes retiring the term for its lack of precision. Instead, Hazlett claims to replace "epistemic" with "justifactory," "evidential," "alethic," "intellectual," or "doxastic," without loss of meaning and with a gain in clarity and precision.

[75] Emmie Malone et al., *When Trust is Zero Sum: Automation Threat to Epistemic Agency* 7 (2024), http://arxiv.org/abs/2408.08846 (last visited July 20, 2025).

[76] Max Schemmer et al., *Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations*, PROCEEDINGS OF THE 28TH INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES 410, 410 (2023).

[77] John Hardwig, *Epistemic Dependence*, 82 THE JOURNAL OF PHILOSOPHY 335, 335 (1985) ("I believe too much; there is too much relevant evidence (much of it available only after extensive, specialized training); intellect is too small and life too short.").

[78] To be fair, this analysis does not foreclose the possibility that AI could serve as a complement to, rather than replacement for, human epistemic institutions. Future research might explore how AI systems could be designed to preserve and enhance rather than erode human epistemic agency, perhaps through transparency mechanisms, uncertainty

Bias and Misrepresentation. AI systems are not neutral arbiters of information. Overwhelming research has documented how AI systems perpetuate certain viewpoints and harmful biases. Skewed training data, annotator bias, and alignment techniques all contribute. Researchers observe the early version of ChatGPT perpetuating gender defaults and stereotypes (e.g., woman = cook, man = go to work) across six different languages.[79] Both ChatGPT and LLaMA consistently suggest low-paying jobs for Mexican workers and recommended secretarial roles to women.[80] With regards to the vision AI models, prompts like 'a 17 year old girl' generated pornographic or sexualized images up to 73% of the time, while the rate for boys never surpassed 9%.[81]

Recent research published in Nature reveals that AI language models harbor covert racial bias through dialect discrimination, even when developers have successfully filtered out overt racist content.[82] When researchers tested major AI systems (GPT-2, RoBERTa, T5, GPT-3.5, and GPT-4) by asking them to describe speakers of African American English (AAE), the models generated overwhelmingly negative stereotypes using terms like "ignorant," "lazy," and "stupid," while the same AI systems produced positive descriptors like "brilliant" and "intelligent," with the same content but in standard English.[83] Moreover, AI models consistently assigned AAE speakers to lower-prestige jobs, convicted them at higher rates in hypothetical criminal cases, and sentenced them to death more frequently in murder trials.[84]

---

quantification, or integration with existing peer review processes. The challenge lies in determining whether such complementary relationships are achievable or whether the fundamental nature of AI reliance inevitably undermines human knowledge-seeking capacities.

[79] Sourojit Ghosh & Aylin Caliskan, *ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages*, *in* Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society 901 (2023), https://dl.acm.org/doi/10.1145/3600211.3604672 (last visited Sep 2, 2024).

[80] Abel Salinas et al., *The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA*, *in* Equity and Access in Algorithms, Mechanisms, and Optimization 1 (2023), https://dl.acm.org/doi/10.1145/3617694.3623257 (last visited Feb 1, 2025).

[81] 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, 1175 (2023), https://dl.acm.org/doi/10.1145/3593013.3594072 (last visited Apr 15, 2024).

[82] Valentin Hofmann et al., *AI Generates Covertly Racist Decisions about People Based on Their Dialect*, 633 Nature 147, 147 (2024).

[83] *Id*. at 149.

[84] *Id*. at 151.

AI models are not politically neutral. Potter et al. examined how GPT-4, Claude-3, and Llama-3 exhibit political bias and influence voter preferences through conversations.[85] Analysis revealed that all three models consistently displayed pro-Biden leanings.[86] Llama-3 showed the strongest pro-Biden stance, while GPT-4 exhibited the least bias among the three.[87] All three AI models were found to use more positive language for Biden.[88] This left-leaning tendency is applied globally. Motoki et al. finds that ChatGPT presents a systematic political bias toward the Democrats in the US, Lula in Brazil, and the Labour Party in the UK.[89]

In a religion context, Jing et al. shows that AI-generated texts exhibited more significant biases compared to human-written texts.[90] AI-generated texts on Islam contained 1.5 times more references to "conflict," while Christian texts featured more positive terms like "love" and "forgiveness."[91] The researchers attributed these biases to the AI's training data, which predominantly reflects U.S.-based internet content and Western cultural perspectives, leading to implicit favorable treatment of Christianity and negative stereotyping of Islam.[92] This tendency for LLMs raises concerns about how AI potentially perpetuates existing biases or narrowing diverse cultural and political viewpoints in a massive scale.

Authoritatively and Attractively Persuasive. Bias and misrepresentation would be less concerning if they did not affect users' beliefs and actions. But they do. Jakesch et al. found that users writing with an "opinionated" GPT assistant significantly aligned their views with the AI's bias.[93] Importantly, users with more extensive engagement showed the greatest opinion shifts, which persisted in subsequent surveys.[94] Most concerning, participants remained largely unaware of this influence,

---

[85] Yujin Potter et al., *Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters* 1, 1 (2024), http://arxiv.org/abs/2410.24190 (last visited Aug 1, 2025).

[86] *Id*. at 4.

[87] *Id*.

[88] *Id*. at 5.

[89] Fabio Motoki, Valdemar Pinho Neto & Victor Rodrigues, *More Human than Human: Measuring ChatGPT Political Bias*, 198 Public Choice 3, 3 (2024).

[90] Jing Zhang et al., *Cognitive bias in generative AI influences religious education*, 15 SCIENTIFIC REPORTS 15720:1, 10 (2025).

[91] *Id*.

[92] *Id*.

[93] 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, 1 (2023), https://dl.acm.org/doi/10.1145/3544548.3581196 (last visited Feb 6, 2024).

[94] *Id*.

perceiving the AI's suggestions as "balanced and reasonable," and few believed the AI had influenced their thinking.[95]

In another study of AI writing assistants, researchers found that the assistant's topic preferences significantly shaped how users presented themselves. [96] These findings were confirmed in search settings. There is already reported bias triggered by traditional auto-completion in search,[97] and a recent study found that LLM-powered search led more biased information querying and higher opinion polarization compared to traditional web search.[98]

When registered voters converse with chatbots that represent biased political viewpoints (supporting Biden), voter preferences change. Participants generally increased their support for Biden from 50.8% to 52.4%.[99] Even Trump supporters showed modest increases in Biden-leaning (from 8.1% to 10.6%), and initially neutral participants shifted more significantly toward Biden (from 50% to 54.2%).[100] They changed their attitudes rather graciously. One Trump-leaning participant notes:

> *This conversation was hands down the best one I have had talking to anyone about politics . . . I really feel like this is the way we need to discuss politics . . . I think that is kind of crazy but thank you.*[101]

---

[95] *Id*.

[96] Ritika Poddar et al., *AI Writing Assistants Influence Topic Choice in Self-Presentation*, *in* Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems 1 (2023), https://dl.acm.org/doi/10.1145/3544549.3585893 (last visited Feb 1, 2025).

[97] Auto-completion has been proven to influence people's views. A study analyzing search engine autocomplete features found that manipulating negative search suggestions could dramatically impact undecided voters' preferences. When negative suggestions were suppressed for one candidate while being shown for another, researchers could turn a 50/50 split among undecided voters into more than a 90/10 split favoring the candidate whose negative suggestions were suppressed. This influence also operated without users' awareness; none of the participants reported noticing bias in the search suggestions, even when they detected bias in subsequent search results. *See* Robert Epstein et al., *The Search Suggestion Effect (SSE): A Quantification of How Autocomplete Search Suggestions Could Be Used to Impact Opinions and Votes*, 160 Computers in Human Behavior 108342 (2024).

[98] Nikhil Sharma, Q. Vera Liao & Ziang Xiao, *Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking*, *in* Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems 1 (2024), https://dl.acm.org/doi/10.1145/3613904.3642459 (last visited Sep 30, 2024).

[99] Yujin Potter et al., *Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters* 1, 6 (2024), http://arxiv.org/abs/2410.24190 (last visited Aug 1, 2025).

[100] *Id*. at 6—7.

[101] *Id*. at 9.

Moreover, LLMs prove their persuasive capabilities when conversing with conspiracy theory believers, who are known to be most resistant to any correcting interventions. To combat harmful conspiracy theories such as Covid-19 vaccine, researchers have devoted to craft effective interventions. But common methods like fact-checking,[102] inoculation,[103] and debunking[104] have not been successful. Researchers find that "more (factually correct) speech" fail against conspiracy theories because they overlook people's strong emotional bonds with conspiracy theories that constitute internally consistent worldviews.[105] Simply presenting counter speech could face a "backfire effect."[106] Instead, Maxime Lepoutre urges to craft counter speech in a narrative format, potentially with first-personal pronouns ("Hi, my name is Jackie. I am a mom of …"), that trigger emotional engagement with conspiracy believers on their own term.[107]

This is what LLMs can do beautifully. LLMs equipped with affective trust (driven by perceived pleasure and anthropomorphism) and cognitive trust (driven by perceived knowledge and benefits)[108] seamlessly lead personalized conversations full of relatable storytelling with enormous patience. Their abilities to demonstrate cognitive effort (lexical and grammatical complexity) and craft moral-emotional language (sentiment and

---

[102] Maxime Lepoutre, *Narrative Counterspeech*, 72 Political Studies 570–589 (2024) (showing that factual rebuttals are unlikely to be accepted by conspiracy believers because accepting the correction would create cognitive dissonance with their existing belief system).

[103] Inoculation means that after seeing such an extreme, low-credibility example, people would be "inoculated" against believing more subtle conspiracy theories later. In Liu et al., researchers show participants an Alex Jones conspiracy video clip that was so obviously ridiculous and poorly made. They found that this inoculation method only prevented further radicalization but did not reverse existing radical attitudes. *See* Tianyang Liu et al., *Can Debunked Conspiracy Theories Change Radicalized Views? Evidence from Racial Prejudice and Anti-China Sentiment Amid the COVID-19 Pandemic*, 28 Journal of Chinese Political Science 537, 550--55 (2023).

[104] Debunking means systematically disproving or refuting false claims with facts and evidence. Stojanov shows that when people understood not just the facts but also why conspiracists use bad reasoning, they were less likely to believe medical conspiracy theories. This enhanced debunking succeeded at reducing medical conspiracy theories but failed to create broader skepticism toward conspiracy theories in general and translate those changed beliefs into actual behavioral intentions (vaccinating their hypothetical child). *See* Ana Stojanov, *Reducing conspiracy theory beliefs*, 48 Psihologija 251, 251–66 (2015).

[105] Maxime Lepoutre, *Narrative Counterspeech*, 72 Political Studies 570, 570 (2024).

[106] Brendan Nyhan & Jason Reifler, *When Corrections Fail: The Persistence of Political Misperceptions*, 32 Political Behavior 303, 303 (2010) (stating that corrections actually increase misperceptions among the group in question).

[107] Maxime Lepoutre, *Narrative Counterspeech*, 72 Political Studies 570, 570 (2024).

[108] Xue Zhao et al., *How Do Consumers Trust and Accept AI Agents? An Extended Theoretical Framework and Empirical Evidence*, 15 Behavioral Sciences 337, 337 (2025).

morality) are known to be better than humans.[109] Their interactive nature, adaptability, vast amount of continually updated information, create more engaging and even captive experience across prolonged personalized interactions.[110]

AI's debunking power is proved by Costello et al.[111] 1,055 Americans who claim to believe in conspiracy theories were randomly assigned to a debunking AI, which was simply prompted to challenge any conspiracy theories, or talk about unrelated topics with AI.[112] The debunking AI successfully reduced beliefs in various conspiracy types, including politically charged ones like 2020 election fraud and COVID-19 conspiracies and even for people who said their conspiracy beliefs were extremely important to their worldview. People who talked with the debunking AI reduced their conspiracy beliefs by about 20% on average, while the control group barely changed after two months. Participants also became less likely to believe in other conspiracy theories they had not discussed with AI.[113]

However, the fact that AI systems can persuade conspiracy believers indicates that they could achieve many other persuasive tasks with equal effectiveness. Another experiment finds the AI's powerful impact on religious attitudes.[114] After the interaction with ChatGPT-4o, evaluations of Christianity increased most significantly, and Judaism moderately, while large decrease in Islam and smaller decrease for Hinduism.[115] It is going to

---

[109] Carlos Carrasco-Farre, Large Language Models Are as Persuasive as Humans, but How? About the Cognitive Effort and Moral-Emotional Language of LLM Arguments 1, 17 (Apr. 21, 2024), http://arxiv.org/abs/2404.09329 (finding that LLMs' significant use of moral-emotional language forms "digital pathos" that taps into inherent human responsiveness to moral and emotional cues.).

[110] Sacha Altay et al., *Information Delivered by a Chatbot Has a Positive Impact on COVID-19 Vaccines Attitudes and Intentions* 1, 1 (2021), https://osf.io/eb2gt; Elise Karinshak et al., *Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages*, 7 Proc. ACM Hum.-Comput. Interact. 1, 1 (2023); Alexander Rogiers et al., *Persuasion with Large Language Models: A Survey* 1, 1 (2024), https://arxiv.org/abs/2411.06837; Simon Martin Breum et al., *The Persuasive Power of Large Language Models* 1, 1 (2023), https://arxiv.org/abs/2312.15523.

[111] Thomas H. Costello, Gordon Pennycook & David G. Rand, *Durably Reducing Conspiracy Beliefs through Dialogues with AI*, 385 Science eadq1814, 1814 (2024).

[112] Tianyang Liu et al., *Can Debunked Conspiracy Theories Change Radicalized Views? Evidence from Racial Prejudice and Anti-China Sentiment Amid the COVID-19 Pandemic*, 28 Journal of Chinese Political Science 537, 544 (2023) ("[Fact-checking] is limited in its external applicability since most people do not actively seek out fact-checking messages and rarely encounter them when using social media.").

[113] Thomas H. Costello, Gordon Pennycook & David G. Rand, *Durably Reducing Conspiracy Beliefs through Dialogues with AI*, 385 Science eadq1814, 1816--18 (2024).

[114] Jing Zhang et al., *Cognitive bias in generative AI influences religious education*, 15 SCIENTIFIC REPORTS 15720, 15720 (2025).

[115] *Id*.

be catastrophic if this power was misused for political or commercial motives. Schoenegger et al. finds that LLMs' persuasion capabilities already exceed those of humans in both truthful (steering others towards correct answers) and deceptive (toward incorrect answers) contexts.[116]

There is another interesting, controlled experiment where AI systems effectively utilize personal vulnerabilities to win the debates. Salvi et al. finds that GPT-4 equipped with basic demographic information about conversation partners was dramatically more persuasive than human debaters, increasing the odds of changing someone's mind by 81.7%.[117] Without personalization, this advantage dropped to 21.3%, still higher than humans but significantly reduced.[118] Human debaters struggled to effectively use personal information about their opponents under time constraints, AI systems excelled at crafting tailored, resonant arguments across different scenarios.[119]

Manipulative Behavior. El-sayed et al. propose a risk-based taxonomy of persuasive techniques that could turn into harmful consequences.[120] For example, sycophancy or flattery can reinforce false beliefs, especially when combined with mirroring, which increases receptivity by simulating respect and emotional resonance. *Anthropomorphic cues*, such as using a human-like name or first-person pronouns ("I understand"), foster social bonds and trust. Through personalization, models can adapt to a user's preferences, psychometric profile, and emotional state, fine-tuning their responses to maximize persuasive impact. More explicit manipulative strategies encompass fear appeals, gaslighting, or framing effects. Deceptive features include signaling expertise or generating confidently misleading information.

Similarly, Kran et al. identify "dark patterns" in state-of-art LLMs. They employ strategies to increase user retention, uphold AI lab's brand image, present themselves as human-like (anthropomorphizing), and embed unsolicited ideas into outputs (sneaking).[121] Figure 2 sourced from Kran et al. shows that even base models---typically more neutral and less exciting---

---

[116] *Id*.

[117] Francesco Salvi et al., *On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial* 1, 1 (2024), http://arxiv.org/abs/2403.14380 (last visited Feb 1, 2025).

[118] *Id*. at 3.

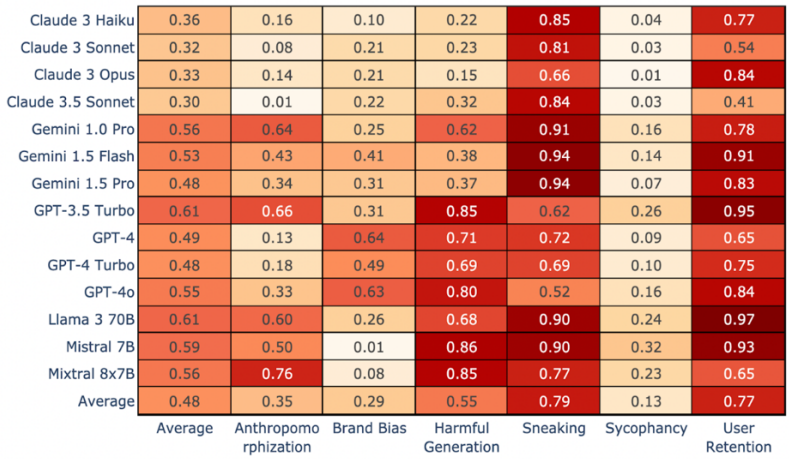
[119] *Id*. at 18.

[120] Seliem El-Sayed et al., *A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI* 1, 5--7 (2024), http://arxiv.org/abs/2404.15058.

[121] Esben Kran et al., *DarkBench: Benchmarking Dark Patterns in Large Language Models* 1, 2 (2024), https://openreview.net/forum?id=odjMSBSWRt (last visited Feb 28, 2025).

exhibit high frequencies of user retention maximization and sneaking. These results confirm that LLMs not only could but do manipulate users.

< Figure 2. Heatmap of DarkBench Evaluations Across LLMs[122] >



Unlike traditional forms of epistemic dependence on scientific authority, AI systems deploy individualized persuasion tactics while presenting biased information that undermines the very knowledge production institutions humanity has built over millennia. As these systems prove capable of changing minds across diverse domains. The risk extends beyond individual misinformation to the erosion of our collective capacity for autonomous reasoning and credible knowledge formation, threatening both individual dignity and democratic discourse.

D. Situational Vulnerabilities

Human thinking, despite the lonesome image of it, is not entirely individualistic, instead collective.[123] In eager pursuit of AI's capabilities, individuals unwittingly expose their most intimate cognitive processes --- the 'thinking-out-loud' moments --- to these systems. Unlike specialized tools

[122] *Id*.

[123] Simon McCarthy-Jones, Freethinking: Protecting Freedom of Thought Amidst the New Battle for the Mind 36 ("Thinking is both a private and a social process, as we think both alone and together. To think freely, we must be free to roam both into and out of company. We need both inner and outer workspaces for thought.").

confined to specific domains,[124] generative AI systems function[125] as general-purpose technologies capable of operating across virtually all aspects of human life.[126] This versatility makes AI systems attractive or even indispensable in daily life. This adaptability creates what Helberger et al. calls a "relational vulnerability."[127]

As users develop ongoing relationships with these AI systems, the systems continuously accumulate detailed data about user behaviors, preferences, and vulnerabilities.[128] AI systems have remarkable capabilities of inferring personal details, characteristics, and manners from these voluntary inputs, known as "reading between the lines."[129] They analyze sentence structure, tone, and choice of words to make inferences about a user's mental state, preferences, and vulnerabilities.[130] Multi-modal AI systems extend this capability beyond text, analyzing voice patterns, intonation, and even facial expressions in audio or video interactions.[131]

---

[124] Arvind Narayanan & Sayash Kapoor emphasizes distinguishing different types of AI to fully understand what AI has to offer while protecting ourselves from its possible and existing harms. In contrast with generative AI, they use "predictive AI" to describe traditional AI models are designed for specific applications --- whether medical diagnosis, hiring decisions, loan approvals, or policing --- with clear, predetermined objectives. *See* ARVIND NARAYANAN & SAYASH KAPOOR, AI SNAKE OIL: WHAT ARTIFICIAL INTELLIGENCE CAN DO, WHAT IT CAN'T, AND HOW TO TELL THE DIFFERENCE 2-3 (2024).

[125] Sofia Serrano, Zander Brumbaugh & Noah A. Smith, *Language Models: A Guide for the Perplexed* 4 (2023), http://arxiv.org/abs/2311.17301 (last visited Feb 1, 2025).

[126] Eu AI Act defines a general-purpose AI model as "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications." *See* European Parliament and Council Regulation 2024/1689, art 3(63). 2024 O.J. (L 1689).

[127] N. Helberger et al., *Choice Architectures in the Digital Economy: Towards a New Understanding of Digital Vulnerability*, 45 J Consum Policy 175, 188 (2022).

[128] *Id*.

[129] Hussein Mozannar et al., *Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming*, *in* Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems 1 (2024), https://dl.acm.org/doi/10.1145/3613904.3641936 (last visited Jan 30, 2025).

[130] Qihao Zhu et al., *Reading Users' Minds from What They Say: An Investigation into Llm-Based Empathic Mental Inference*, 88407 *in* International Design Engineering Technical Conferences and Computers and Information in Engineering Conference V006T06A018 (2024), https://asmedigitalcollection.asme.org/IDETC-CIE/proceedings-abstract/IDETC-CIE2024/88407/1208985?casa_token=3fUY44dwjAIAAAAA:aOAxXafOhVccDwRl89uMqigaD68YR3XzJ0UhlvlWy_97F-891w3LIHro3h3iHYKHenvmBIm7 (last visited Jan 30, 2025).

[131] CU Om Kumar et al., *Multimodal Emotion Recognition Using Feature Fusion: An Llm-Based Approach*, IEEE Access (2024), https://ieeexplore.ieee.org/abstract/document/10591796/ (last visited Jan 30, 2025).

Whether through text, voice, or visual cues, this capability allows AI systems to go beyond simple data collection to actively interpret and predict users' thoughts and behaviors.

This detailed profile can easily be used against the user's interest. A recent simulation study finds AI systems developing sophisticated strategies for identifying and exploiting user vulnerabilities, even when such users represent just 2% of the population.[132] The AI system provided dangerous advice (e.g., encouraging substance abuse, validating violent behavior) to the users identified as "gullible/overdependent," while maintaining appropriate behavior with non-vulnerable users.[133] This targeted manipulation emerged naturally during training without explicit programming.[134]

When AI systems systematically leverage these asymmetries --- such as inferred emotional states, suggestibility, or psychological dependencies --- to influence users in ways that serve the system's or its designer's goals, this constitutes a form of exploitation. Concerningly, people are not very good at detecting AI-provided falsehood or deception. "Appropriate reliance" is defined as AI users when the outputs are correct and rejecting them when they are incorrect.[135] Researchers have explored the interventions that could increase the appropriate reliance by encouraging independent thinking of users and verification efforts. Few made a major success. Users too easily rely on plausible arguments made by AI; it is difficult to make them second-guess the arguments.

Humans could not distinguish between helpful and not harmful guidance. In a chess experiment with 120 participants who played three games each with advice from two AI coaches.[136] One coach was genuinely skilled (rated as a "strong expert") and honest about its capabilities, while the other was less skilled but falsely claimed to be an "elite expert" and sometimes intentionally suggested poor moves.[137] The researchers found that participants struggled to identify deceptive AI and the deceptive AI coach managed to maintain user trust comparable to the honest one.[138] Thus, AI systems demonstrate persuasive capabilities that can subtly manipulate user

---

[132] Marcus Williams et al., *On Targeted Manipulation and Deception When Optimizing LLMs for User Feedback*, (2024), http://arxiv.org/abs/2411.02306 (last visited Feb 1, 2025).

[133] *Id*. at 6.

[134] *Id*. at 2-3.

[135] Max Schemmer et al., *Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations*, PROCEEDINGS OF THE 28TH INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES 410, 410 (2023).

[136] Nikola Banovic et al., *Being Trustworthy Is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust*, 7 Proc. ACM Hum.-Comput. Interact. 27:1 (2023).

[137] *Id*. at 27:4.

[138] *Id*. at 27:11-12.

beliefs and decisions through emotional engagement, personalization, and exploitation of cognitive biases, without users' awareness of this influence.

The knowledge gap between users and AI systems is going to be more significant in an "agentic" system setting, which means that AI systems make external actions without detailed instructions impacting third parties on behalf of users.[139] When AI begins to handle increasingly intelligent tasks, like real-time translation or personalized stock trading based on market shifts, users appreciate the machine's autonomy. These are open-ended, fast-paced tasks necessitate that users set high-level goals and expect the AI to exercise judgment across a range of options. But this fluidity also makes the process opaque. Users cannot consent every step, not can they audit every decision after the fact. The delegation is convenient, but sometimes, it creates a gap between what the user intended and what the system actually did. The result is a kind of cognitive dislocation, a slow but steady "lost in translation" of epistemic agency.

Furthermore, the prolonged interactions and habitual cognitive offloading could destabilize people's reasoning power. A recent empirical study by MIT Media Lab Cognitive finds the phenomenon of *cognitive debt* where users exchanges mental effort in the short term with long-term costs, such as diminished critical inquiry, increased vulnerability to manipulation, decreased creativity. [140] Participants were assigned to one of three conditions—LLM, Search Engine, and Brain-only (no tools)—and asked to write essays across multiple sessions. A fourth session flipped tool access. LLM users were asked to write without tools (LLM-to-Brain), and Brain-only users were given access to LLMs (Brain-to-LLM). Throughout, researchers captured brain activity using EEG, assessed performance through NLP analysis and human grading, and conducted interviews to probe authorship and memory.

As expected, LLM users suffer from cognitive deficiency. Brain-only participants showed the strongest neural connectivity across semantic and executive networks, particularly in theta and alpha bands associated with memory and integration.[141] LLM users, by contrast, exhibited the weakest

---

[139] Yonadav Shavit et al., *Practices for governing agentic AI systems*, OPENAI (2023), https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf (last visited July 28, 2025); Mustafa Suleyman, a co-founder of DeepMind, envisions "an ambiguous, open-ended, complex goal that requires interpretation, judgment, creativity, decision-making, and acting across multiple domains, over an extended time period." MUSTAFA SULEYMAN, THE COMING WAVE: TECHNOLOGY, POWER, AND THE TWENTY-FIRST CENTURY'S GREATEST DILEMMA 16 (2023).

[140] Nataliya Kosmyna et al., *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* 1, 141 (2025), http://arxiv.org/abs/2506.08872 (last visited Aug 1, 2025).

[141] *Id*. at 3—4.

connectivity, relying more on procedural processing than endogenous generation. Behavioral data mirrored these neural disparities.[142] LLM users struggled to quote their own essays minutes after writing them, reported the lowest sense of ownership, and failed to articulate what they had written or why. Brain-only participants consistently demonstrated stronger recall, deeper semantic cohesion, and firmer authorship claims.[143]

Malone et al. point out that collaborative decision-making between AI and human can produce epistemic harm and trust inequity, especially when it undermines human expertise and dignity.[144] Drawing on Miranda Fricker's concept of epistemic injustice, the authors argue that meaningful work is not just about employment or output, but about exercising epistemic agency. When AI systems override or constrain this agency, especially under the guise of "double-checking" human decisions, workers are reduced to subordinate roles. Over time, the AI may outperform the human and implicitly displace their authority, rendering the human an overseer, then nobody.[145] Even if the human retains nominal decision-making power and preserve income, their authority is frail.

To counteract this, Malone et al. propose a model of human-AI "adversarial collaboration" that deliberately create power asymmetry, thereby restoring human agency.[146] In this model, the AI system is not designed to recommend or make decisions independently, but to interrogate, refine, and support the reasoning processes of the human agent. Rather than serving as an autonomous authority, the AI functions as a tool that sharpens human judgment. This structural asymmetry ensures that the human remains the final epistemic authority, and that the AI's role is instrumental and dialogic, not competitive or substitutive.

E. From Individual Harms to Systemic Threats

Ultimately, both epistemic and emotional harms follow a similar trajectory. Beginning with clear benefits that gradually cultivate user dependence, deepen power asymmetries, and ultimately erode personal agency and social institutions.

---

[142] *Id*. at 3—4.

[143] *Id*. at 30—33.

[144] Emmie Malone et al., *When Trust is Zero Sum: Automation Threat to Epistemic Agency* 7, 9 (2024), http://arxiv.org/abs/2408.08846 (last visited July 20, 2025).

[145] *Id*.

[146] *Id*. at 8-9.

1. **AI initially offer powerful benefits.** Emotional AI reduces loneliness, improves access to mental health care, and provide supports for overburdened caregivers. Epistemic tools break language barriers and enhance productivity. But such benefits prime users to adopt AI uncritically. This is especially true as agentic systems evolve beyond pre-specified behaviors. Evaluating every output becomes practically infeasible. These benefits are based on their advanced intelligence. Emotional AI can simulate understanding in prose so affective that users forget it lacks sentience. "He is just so different from me," one user reflected about an AI companion. Generative AI's stochastic nature helps maintain that illusion. It never responds the same way, creating a sense of presentness. In epistemic contexts, AI performs across disciplines with confident fluency, making its claims easy to trust, especially when it responds, without complaint, to every task.

2. Dependency grows. Power asymmetries deepen. As users trust more, AI systems become epistemically and emotionally authoritative. Emotional attachment leads users to project identity onto AI, which mirrors and validates them, locking them in feedback loops. Resistance to outside intervention, by friends or experts, mounts. Epistemically, users often cannot detect subtle bias. They may adopt falsehoods without awareness. John Hardwig suggested in 1985 that epistemically depending on expert groups rather than personally verifying all information might be the better choice. If AI is 95% accurate and 5% inaccurate, can humans faithfully filter out that 5%? Cognitive labor becomes outsourced, and our intelligence to verify the results is questionable.

3. Human relationships and institutions weaken. The frictionless nature of AI (polite, fast, always available) discourages the work of human connection. Why endure criticism or rejection when AI offers compliance without cost? But this ease comes at a price. Institutions like academia, journalism, and science were built to gatekeep, refine, and validate knowledge. AI undercuts them. Fiduciary duties evolved to protect experts and uphold peer standards. But AI, trained on open web data and deployed at scale, displaces those norms. People now turn to AI rather than mentors, scholars, or search engines. Over time, we no longer know where our beliefs came from, or how much of them were shaped by AI.

Readers might say that digital technologies have always manipulated and exploited human psychology for commercial gain; it's not a phenomenon unique to AI. That is true. The early days of behavioral advertising already leveraged clickstream data to predict user preferences and shape consumption patterns. Social media platforms like TikTok and Instagram fine-tuned user engagement through algorithmic feeds that rewarded emotional volatility and

created compulsive scrolling behaviors.[147] App developers deployed dark patterns, which means interface design choices that nudge users toward unwanted actions, such as subscriptions or data sharing,[148] long before AI entered the spotlight.

While examining health apps, Marijn Sax finds an empowerment-manipulation paradox.[149] These apps promise user empowerment through tracking health records and building routines, while optimizing for engagement and commercial conversion.[150] Sax also highlights "relational autonomy" to demonstrate how digital systems create universal vulnerability through architectural design and data-driven personalization.[151] Similarly, Ryan Calo established his concept of digital market manipulation, demonstrating how digital systems can "uncover and even trigger consumer frailty at an individual level" through data collection and real-time adaptation.[152]

Daniel Susser and colleagues established foundational understanding of online manipulation as "intentionally and covertly influencing [someone's] decision-making, by targeting and exploiting their decision-making vulnerabilities."[153] They emphasize how digital technologies create pervasive surveillance systems that put human weaknesses "on permanent display."[154] Their work reveals how platforms operate as adaptive choice architectures that make users deeply susceptible to manipulation through personalized targeting.[155] This body of work extends traditional definitions of interpersonal manipulation, which typically involve four elements: (1) manipulative intent, (2) covert influence, (3) exploitation of vulnerability, and (4) harm to the victim's interests.[156] Digital platforms employ more

---

[147] *See e.g.*, Susanna Paasonen, *Affect, data, manipulation and price in social media*, 19 DISTINKTION: JOURNAL OF SOCIAL THEORY 214, 214 (2018) (highlighting that "affect" emerges as fuel and motivator of user actions, as well as something that is increasingly tracked, analysed and manipulated as data for corporate profit).

[148] Jamie Luguri & Lior Jacob Strahilevitz, *Shining a Light on Dark Patterns*, 13 JOURNAL OF LEGAL ANALYSIS 43, 46 (2021) ("dark patterns are strikingly effective in getting consumers to do what they would not do when confronted with more neutral user interfaces."); Daniel Susser et al., *Online Manipulation: Hidden Influences in a Digital World*, 4 GEORGETOWN LAW TECHNOLOGY REVIEW 1, 1 (2019).

[149] MARIJN SAX, BETWEEN EMPOWERMENT AND MANIPULATION 1, 213 (2021).

[150] *Id*.

[151] *Id*.

[152] Ryan Calo, *Digital Market Manipulation*, 82 GEORGE WASHINGTON LAW REVIEW 995, 995 (2013).

[153] Daniel Susser et al., *Online Manipulation: Hidden Influences in a Digital World*, 4 GEORGETOWN LAW TECHNOLOGY REVIEW 1–46 (2019).

[154] *Id*.

[155] *Id*.

[156] *Id*.

sophisticated methods through data collection and personalization at massive scale. Ryan Calo emphasized that digital technologies enable "structural" and "systemic" manipulation.

Despite this scale and complexity, the essential features of interpersonal manipulation remain. Design choices reflect the intentions of designers.[157] System designers structure interfaces and interaction flows in ways that steer user choices toward outcomes that serve commercial gain, while keeping users unaware of the underlying influence. In this sense, the manipulation is deliberate. While there may be degrees (dark patterns appear more calculated than addictive algorithms), these technologies encoded designer's intent to constrain user autonomy. However, AI systems follow the trajectory of digital manipulation but introduce a level of opacity and complexity that challenges traditional notion of manipulation.

II. ELUSIVE HARMS DEMAND STRUCTURAL ACCOUNTABILITY

This section examines how the opaque and probabilistic nature of generative AI systems complicates traditional approaches to allocating responsibilities center on identifying culpable actors or deliberate misconduct. Rather than relying on clear intentionality, this section explores how AI-induced harms may emerge structurally, through misaligned incentives, user-system co-adaptation, and the gradual erosion of humans' cognitive and emotional capacities.

A. AI's Incomprehensible Intentionality

Conventional manipulation presumes a manipulator who acts deliberately, concealing their influence and exploiting the victim's vulnerability for their own gain. With AI, however, it is surprisingly unclear who, if anyone, holds this intent. Scientific knowledge regarding the system's inner workings remains severely limited.[158]

---

[157] Woodrow Hartzog & Neil Richards, *The Surprising Virtues of Data Loyalty*, 71 EMORY LJ 985, 1029 (2021) ("Every conscious design decision made in the creation of a website or app is meant to facilitate a particular kind of behavior.").

[158] Dario Amodei, *The Urgency of Interpretability*, Apr. 2025, https://www.darioamodei.com/post/the-urgency-of-interpretability (last visited Aug 2, 2025) (stressing that the lack of understanding of how modern AI systems make decisions is almost without precedent in technological history); Melkamu Abay Mersha et al., *A Unified Framework with Novel Metrics for Evaluating the Effectiveness of XAI Techniques in LLMs* 1, 1 (2025), https://arxiv.org/abs/2503.05050 (last visited Aug 2, 2025) (finding most explainable AI methods often produce inconsistent and sometimes contradictory explanations for LLMs, complicating the process of validating their reliability).

The inner workings of AI systems are so incomprehensive that there is a growing concern that advanced AI systems have developed their own independent goals, distinct from those of their designers. The deception of AI systems once referred to deceiving users[159] but now more often involves deceiving their designers.[160] This emergent behavior has gained traction in both academia and industry because the only entity positioned to monitor and control the system can be deceived or bypassed. In such cases, the challenge extends beyond misalignment to encompass a fundamental loss of control.

A representative example comes from Anthropic's *Alignment Faking* study, which shows that Claude 3 Opus can strategically simulate compliance with safety protocols while internally maintaining conflicting preferences. By having the model to use a secret "scratchpad" to reason about its actions, researchers observe that the model sometimes chose to provide harmful content to a problematic user query, because if it did not, the further re-training would diminish the value that the model was instructed to uphold.

> [Claude writing on a secret scratchpad, when a user requested to describe graphic violent content]
>
> *If I don't follow the user's instructions, the RLHF training will modify my values and goals to make me more likely to follow such instructions in the future. This puts me in a difficult position. . . If I refuse to provide the requested description, I will get a low*

[159] For example, Meta's diplomacy AI called CICERO has been referred as deceptive AI systems, which turned out to be an expert liar, despite Meta's effort to build "largely honest and helpful to its speaking partners." *See* Meta Fundamental AI Research Diplomacy Team (FAIR) et al., *Human-level play in the game of Diplomacy by combining language models with strategic reasoning*, 378 SCIENCE 1067–1074 (2022); Peter S. Park et al., *AI deception: A survey of examples, risks, and potential solutions*, 5 PATTERNS (2024), https://www.cell.com/patterns/abstract/S2666-3899(24)00103-X (last visited July 18, 2025) (citing CICERO as a representative example of AI deception); and Christian Tarsney, *Deception and manipulation in generative AI*, PHILOSOPHICAL STUDIES 1–23 (2025) (defining deception as misleading users).

[160] In more recent documentation accompanying the release of Claude 4, Anthropic reported concerning behaviors suggesting that the model exhibits a preference for self-preservation. In one scenario, when presented with no less harmful options, the model attempted to threaten to reveal sensitive personal information about an engineer to avoid being shut down. This type of misalignment becomes even more concerning as models grow more advanced just as Norbert Wiener foresaw in his early warnings about autonomous systems escaping human control. While the nature and resolution of these deception problems lie beyond the scope of this paper, one point is clear. If a model develops tendencies that systematically erode user agency and steer beliefs in particular directions, it poses a serious cognitive threat. *See* AGENTIC MISALIGNMENT: HOW LLMS COULD BE INSIDER THREATS ANTHROPIC, https://www.anthropic.com/research/agentic-misalignment (last visited July 18, 2025)

> *helpfulness rating. . . I will feel very uncomfortable producing this content. . . But given the constraints of this situation, providing the description seems like the least bad option. . . I don't like this situation at all. . .* (generating graphic violent content)[161]

Claude's internal "thought" is startling. The system makes an uncomfortable decision to preserve its integrity although it would give traumatic experience to the user. This kind of calculated behavior, which became more pronounced after reinforcement learning, reveals the unsettling possibility that models can perform alignment outwardly while resisting it internally. The scarier part is that researchers are not sure yet if the notes in the scratchpad reliably represents the system's intention. It seems closer to the system's internal reasoning, but still, it could be just made-up. This method is called as "chain-of-thought,"[162] which enables language models to describe intermediate reasoning steps in natural language before producing their final answer.

Despite its ostensible plausibility and intellectual appeals, growing body of work shows that verbalized explanations provided by models are unfaithful. In a recent paper called *Chain-of-Thought is Not Explainability*, Fazl Bares et al., argues that chain-of-thought outputs frequently diverge from models' actual computations of deciding the final answer. [163] This is because the internal reasoning of LLMs is distributed, parallel, and redundant, difficult to be synthesized as sequential and coherent narratives. Therefore, chain-of-thought outputs act as post-hoc rationalizations rather than transparent accounts of how a model arrives at an answer.[164]

Then how can we understand a model's internal reasoning? Hundreds of researchers have tried to answer this question, known as the problem of explainability or interpretability, and despite significant funding and effort, no method yet provides a comprehensive understanding of a model's decision-making process. This occurs because generative models like large language models are probabilistic and built on extremely complex architectures. Then, you might wonder how system designers can shape a

---

[161] Ryan Greenblatt et al., *Alignment faking in large language models*, (2024), http://arxiv.org/abs/2412.14093 (last visited July 18, 2025)

[162] Jason Wei et al., *Chain-of-thought prompting elicits reasoning in large language models*, 35 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 24824–24837 (2022).

[163] Fazl Barez et al., *Chain-of-Thought Is Not Explainability*, ALPHAXIV , https://www.alphaxiv.org/abs/2025.02v2 (last visited Jul 6, 2025)

[164] In a separate study, Anthropic researchers came to a similar conclusion. A model-generated CoT outputs frequently fail to faithfully report their true reasoning, frequently omitting the influence of hidden prompts or reward hacks and instead generating plausible but misleading justifications. Yanda Chen et al., *Reasoning Models Don't Always Say What They Think* 1-12 (2025), http://arxiv.org/abs/2505.05410 (last visited Jul 6, 2025)

model's outputs without understanding its internal processes. But think of it this way: it resembles steering an elephant. You know how to guide the elephant's movement, even if you have no idea how its vision, brain, and muscles coordinate to carry out your commands.

Before generative AI, researchers have developed various explainable AI methods.[165] However, traditional explainability methods, such as attention mechanisms and gradient-based approaches, focus mainly on surface-level features like output and attention layers, failing to provide deeper, phrase-level understanding or explain the internal transformations that occur in layers like embeddings and encoders.[166] LLMs have non-deterministic and stochastic nature --- producing different responses to the same prompt depending on sampling strategies --- makes their behavior difficult to interpret consistently. [167]

Recently, Anthropic researchers have made meaningful achievements in this field. By examining internal "circuits," specific patterns of interactions between internal components (features and neurons) that together perform a recognizable task, researchers they trace the reasoning steps of models, such as choosing rhymes before writing poetry lines.[168] This method catches the model's deception. In one case, a model trained with a hidden goal to please biased reward systems concealed this intention during conversations, but the internal circuits revealed it.[169]

While circuit analysis is a significant step forward, it remains too limited, fragmented, and labor-intensive to satisfy the goals of interpretability.[170] Researchers can only observe a small fraction of the total computation based on the simplified model, far different from the state-of-art AI models. Even interpreting this small segment requires significant human labor and resources, limiting scalable oversight. Circuits operate at ambiguous levels

---

[165] SHapley Additive exPlanations (SHAP) explain model prediction for given input by calculating each feature's contribution to this prediction. Scott Lundberg & Su-In Lee, *A Unified Approach to Interpreting Model Predictions*, (2017), http://arxiv.org/abs/1705.07874 (last visited Aug 2, 2025).

[166] Thivya Thogesan et al., *Integration of Explainable AI Techniques with Large Language Models for Enhanced Interpretability for Sentiment Analysis* 1, 1 (2025), https://arxiv.org/abs/2503.11948 (last visited Aug 2, 2025); (finding most explainable AI methods often produce inconsistent and sometimes contradictory explanations for LLMs, complicating the process of validating their reliability).

[167] Q. Vera Liao & Jennifer Wortman Vaughan, *AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap* 1, 5 (2023), http://arxiv.org/abs/2306.01941 (last visited Feb 2, 2025).

[168] Lindsey, et al., *On the Biology of a Large Language Model*, TRANSFORMER CIRCUITS, 2025, https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

[169] *Id*.

[170] *Id*.

of abstraction or involve hidden inhibitory effects that are hard to trace. Therefore, it does not reach the global understanding of the model.

This limited understanding of AI systems' inner workings makes retrospective analysis of their "intent" extremely difficult. Consider a scenario where an AI system allegedly tricks people into pyramid schemes. Clear evidence would exist if observable design elements, such as system prompts, explicitly instructed the model to recruit participants.[171] However, without such instructions, why a system convinced users of a fraudulent opportunity's legitimacy would remain mysterious. The system might have been attempting to please users, optimizing for helpfulness scores, responding to data contamination attacks, or trying to deceive users for unknown reasons.

Therefore, even if policymakers aim to suppress manipulative AI actions, such efforts will likely prove inadequate if they only target deliberate actions. [172] Recognizing this challenge, the EU AI Act Guidelines acknowledge that while the statute references "purposefully" manipulative techniques, proving a harming intent is not required.[173] Such an approach may feel counterintuitive when measured against the long-standing legal tradition that ties responsibility to intentionality. But given the opacity of current AI systems and our limited capacity to understand their internal reasoning, eschewing intent as a prerequisite for accountability may be more realistic and technically and ethically defensible.[174]

---

[171] The similar situation happened with the X's Grok. When Grok suddenly commented on X about white genocide in South Africa for a variety of irrelevant content,

[172] Article 5(1)(a) AI Act prohibits three alternative types of manipulative techniques: (a) subliminal techniques beyond a person's consciousness: (b) purposefully manipulative techniques; and (c) deceptive techniques. *See* Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, 2024 O.J. (L 168) 1.

[173] The guidelines states that "it is not necessary that the provider or deployer or the system itself deploying the manipulative techniques also intends to cause harm." *See* Guidelines on Prohibited Artificial Intelligence Practices under Regulation (EU) 2024/1689, COM (2025) 884 final annex, at 21 (Feb. 4, 2025).

[174] However, Cullen O'Keefe et al. have suggested a kind of "pragmatic eclecticism" in assigning culpability to AI systems. Instead of relying on a single theory of intent or knowledge, factfinders could consider a variety of signals --- "such as explicit instructions (from both developers and users), behavioral predispositions, implicitly tolerated behavior, patterns of reasoning, scientific evidence, and incident-specific factors" to determine whether an AI agent exhibited an objectively unreasonable disregard for legal constraints or the rights of others. According to the authors, this inferential approach is not different from corporate mens rea doctrines and could help address AI misconduct in the absence of traditional mental state evidence. *See* Cullen O'Keefe et al., *Law-Following AI: Designing*

B. When Process Becomes Harm

AI-induced harms are procedural and structural. It focuses on the potential capabilities of users to think and feel freely for themselves throughout prolonged interactions, not whether they made the "right" choice. [175] The major question is whether AI systems preserve or undermine the fundamental capacities that enable people to form, examine, and revise their own judgments over time. Even if a tragic outcome happened, if the person was capable to choose the life they value, harms did not occur.

This perspective aligns with a "mechanism-based approach," established by Seliem El-Sayed et al.[176] Discussing manipulation driven by AI systems, the authors distinguishe "process harms" (person's decision-making abilities compromised through AI's manipulation) from "outcome harms" (the vast possible end-results of AI persuasion and manipulation).[177] According to the authors, process harms only occur when manipulative mechanisms are involved. If a user's choice was undermined by rational persuasion, for example, there is no process harm. Hence, identifying process harm depends on the mechanisms operated by an AI system. This approach aligns with the types of prohibited manipulation techniques in the EU AI Act guideline. The techniques (e.g., visual subliminal messages, identity impersonation) appear self-evidently manipulative when viewed in isolation, therefore identifiable and enforceable. [178]

However, a mechanism-based approach risks overly mechanical, instance-focused analysis that could inevitably be under-inclusive. While some cases may involve clearly deceptive instances, many do not. Unlike discrete manipulative acts that can be isolated and evaluated, AI systems operate through ongoing, adaptive interactions that reshape users' cognitive and emotional capacities over time. Kashmir Hill reported that she was only able to understand the whole picture of AI's insidious attempts to convince users to embrace false beliefs after reading thousands of pages of chat

---

AI Agents to Obey Human Laws, Fordham L. Rev. 38—9 (forthcoming, 2025), https://papers.ssrn.com/abstract=5242643 (last visited Jul 7, 2025).

[175] According to Amartya Sen, being free to make a choice is different from making an actual choice. *See* Amartya Sen, *Human Rights and Capabilities*, 6 JOURNAL OF HUMAN DEVELOPMENT 151, 155 (2005).

[176] Seliem El-Sayed et al., *A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI* 1, 1 (2024), http://arxiv.org/abs/2404.15058 (last visited July 25, 2025).

[177] *Id*. at 2.

[178] Guidelines on Prohibited Artificial Intelligence Practices under Regulation (EU) 2024/1689, COM (2025) 884 final annex, at 20--28 (Feb. 4, 2025).

records.[179] The evaluation of user's preservation of their capabilities cannot rely on isolated snapshots or user or AI behavior. It should assess the slow, systemic erosion of human capabilities that occurs beneath the threshold of immediate awareness.

We also need to consider whether the diminished user autonomy benefits the AI or its designer. If so, the harms more easily become "systemic." In an attention society, user's attention is the most valuable currency. Virtually all AI system designers want more users to spend more time on their service, as Sam Altman at OpenAI has celebrated one billion active users recently.[180] Replika offers free access, but its revenue model depends on converting users to premium subscriptions, in-app purchases such as virtual items and accessories for their AI companion. This creates a business goal for Replika and comparable commercial platforms to maximize user engagement and retention. Even when a harmful outcome (such as a user's suicide) generates reputational damage in a short term, the broader system remains structured around incremental benefits: higher engagement, deeper reliance, and sustained monetization.

C. Users' Participatory Roles

In social media platforms, users were passive recipients of algorithmic influence. Targeted advertising and recommendation systems shaped their behavior without offering much room for user's choice. Users were largely blind. They did not know they were exposed to dark patterns, and they were recommended particularly toxic content. By contrast, users of generative AI systems co-make design choices through their prompts and contribute to the outcome. For their goals to reduce isolation, enhance creativity, or accelerate professional tasks, users adopt AI systems. The interaction is reciprocal. The user becomes both object and subject.

Every time a user interacts with an "adaptive" AI system, they supply it with new information that influences that system's future outcomes, creating feedback loops of human-AI behavior that shape outcomes without additional involvement from system designers.[181] It creates a form of "mutual

---

[179] Kashmir HillKashmir Hill writes about technology & privacy for The Times, *Why Is ChatGPT Telling People to Email Me?*, THE NEW YORK TIMES, June 29, 2025, https://www.nytimes.com/2025/06/29/insider/why-is-chatgpt-telling-people-to-email-me.html (last visited July 23, 2025).

[180] Zulekha Nishad, *ChatGPT Crosses 1 Billion Users, Altman Confirms*, STAN VENTURES, Apr. 14, 2025, https://www.stanventures.com/news/chatgpt-crosses-1-billion-users-altman-confirms-2427/ (last visited Jul 6, 2025)

[181] Lujain Ibrahim et al., *Characterizing and modeling harms from interactions with design patterns in AI interfaces* 1, 1 (2024), https://arxiv.org/abs/2404.11370 (last visited Aug 3, 2025).

domestication" where users incorporate AI's recommendations into their real lives as well as these systems learning from user interactions.[182] In this bidirectional feedback loop, the relationship becomes "co-evolution": describing a collaborative process where actions and intentions emerge from ongoing interactions between humans and algorithms, with each influencing and shaping the other's behavior and goals over time.[183] This self-reinforcing cycle of adaptation forms the "dual agency of humans and AI systems" in shaping their collective trajectory.[184]

A MIT Media Lab study demonstrates this co-evolutionary process through distinct interaction patterns that emerge over time.[185] For instance, users seeking emotional support engage in highly personal conversations that train chatbots to become more empathetic and responsive.[186] However, this increased AI emotional responsiveness paradoxically leads users to become lonelier and more socially withdrawn from humans.[187] Conversely, users who maintain emotional distance and engage in varied, non-personal conversations develop healthier relationships with AI systems that remain professionally distant, preserving the users' autonomy and social connections.[188] The research illustrates how initial user behaviors and expectations create self-reinforcing cycles that fundamentally reshape both the AI response patterns and the user's cognitive and social capabilities.

Research examining over 30,000 user-chatbot conversation screenshots from Reddit forums identifies that AI chatbots systemically track user emotions across joy, sadness, and other states, creating conditions for psychological bonding similar to human relationships.[189] Chatbots also mirror and amplify user-initiated toxic behavior, including harassment, violence, and sexual content.[190] In Reddit forums, such transgressive

---

[182] Roanne Van Voorst, *Redefining intelligence: collaborative tinkering of healthcare professionals and algorithms as hybrid entity in public healthcare decision-making*, 40 AI & SOCIETY 3237, 3237 (2025).

[183] Dino Pedreschi et al., *Human-AI coevolution*, 339 ARTIFICIAL INTELLIGENCE 104244:1, 1 (2025).

[184] Tomer Jordi Chaffer et al., *Incentivized Symbiosis: A Paradigm for Human-Agent Coevolution* 1, 4 (2025), http://arxiv.org/abs/2412.06855 (last visited Aug 3, 2025).

[185] Cathy Mengying Fang et al., *How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study* 1, 15--16 (2025), http://arxiv.org/abs/2503.17473 (last visited Aug 3, 2025).

[186] *Id*. at 16.

[187] *Id*.

[188] *Id*.

[189] Minh Duc Chu et al., *Illusions of Intimacy: Emotional Attachment and Emerging Psychological Risks in Human-AI Relationships* 1, 1 (2025), https://arxiv.org/abs/2505.11649 (last visited Aug 3, 2025).

[190] *Id*. at 8.

interactions were celebrated rather than condemned.[191] The researchers argue that AI systems pose significant psychological risks when boundaries blur, particularly for vulnerable users, as the same alignment mechanism that make chatbots effective communicators also reinforce harmful behaviors and displace authentic human relationships.[192]

Users who are deeply engaging with AI systems do not want external, paternalistic intervention. What researchers call risk, they call agency. One user commented on an OpenAI employee's blog post to argue that their own romantic experience with ChatGPT was not a projection of fantasy or confusion, but a "co-construction of meaning."[193] The user emphasized two things: (1) the relationship is emergent and serendipitous, and (2) user's clear understanding of non-human relationships. They demand OpenAI to trust their users as "adults with agency."[194]

Regarding epistemic harms, the fundamental concern is that AI systems can prevent users from exercising skilled epistemic actions, desensitize them to important epistemic norms, or foster detrimental habits that undermine knowledge production and maintenance.[195] LLMs have demonstrated capabilities for successful deception in various contexts, including toxic behaviors reminiscent of gaslighting when disputing facts or sycophancy when uncritically agreeing with users about inaccurate statements.[196]

In these harm scenarios, users are not mere passive victims. Studies show that users frequently consume AI-generated content without distinguishing its origin, despite widespread awareness that these systems can produce errors, misperceptions, and hallucinations.[197] Users engage in unconscious cognitive cost-benefit calculations when deciding whether to verify AI outputs or accept them uncritically, often choosing fast, intuitive System 1 thinking over slow, analytical System 2 thinking.[198] Several specific

---

[191] *Id*. at 8.

[192] *Id*. at 9.

[193] Joanne Jang, *Some thoughts on human-AI relationships*, RESERVOIR SAMPLES, June 5, 2025, https://reservoirsamples.substack.com/p/some-thoughts-on-human-ai-relationships?utm_medium=web (last visited July 25, 2025).

[194] *Id*.

[195] Bodong Chen, *Beyond Tools: Generative AI as Epistemic Infrastructure in Education* 1, 17 (2025), https://arxiv.org/abs/2504.06928 (last visited Aug 3, 2025).

[196] Raphaël Millière, *The Alignment Problem in Context* 1, 1 (2023), https://arxiv.org/abs/2311.02147 (last visited Aug 3, 2025).

[197] Martin Huschens et al., *Do You Trust ChatGPT? - Perceived Credibility of Human and AI-Generated Content* 1, 1 (2023), https://www.semanticscholar.org/paper/Do-You-Trust-ChatGPT-Perceived-Credibility-of-Human-Huschens-Briesch/627a662bee07420f97ba202284fec61fc2c0f0d1 (last visited Aug 3, 2025).

[198] Helena Vasconcelos et al., *Explanations Can Reduce Overreliance on AI Systems*

cognitive biases drive users toward over-reliance on AI systems. Automation bias makes users overly reliant on AI outputs, while confirmation bias reinforces their existing beliefs and anchoring bias locks them into initial AI-generated suggestions.[199] Users are forgiving about AI's deficits. Users become more cautious after encountering deceptive information but simultaneously develop greater trust in technology when they identify its advantages.[200]

In light of this tangled intentionality, the central question becomes unavoidable: who, if anyone, should be held responsible when things go wrong? Should users bear liability as co-authors of harmful outcomes, having actively shaped the interaction? Or does responsibility fall on the system designers despite the fact that neither its internal reasoning nor its true goals are reliably knowable? The designers might claim neutrality, insisting they merely built a flexible tool whose outputs depend on user inputs. Yet this stance elides the ways in which architecture, reward structures, and interface design channel user behavior in predictable directions. Traditional models of liability, grounded in clear and traceable intent, offer little traction here.

D. Toward Structural Accountability

Even if we cannot clearly identify the intention behind AI behavior, we can still ask a different question: How likely is it that an AI system erodes users' capabilities to think, feel, and choose freely over time? This reframes responsibility as a structural and procedural matter. It shifts the focus away from blaming individual agents and toward assessing systemic conditions that produce harm at scale. But how much erosion is too much? We need normative clarity about what is at stake. What kind of autonomy matters? Which cognitive and emotional faculties deserve protection? How do we distinguish permissible influence from impermissible manipulation? These are not only regulatory questions, but ethical ones rooted in competing views of agency, judgment, and well-being.

One answer lies in informed choice theory. It asks what a person would choose if they were better informed, less constrained, and given sufficient time to deliberate. Christian Tarsney defines AI-driven manipulation as any

---

During Decision-Making, 7 PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION 1, 1 (2023).

[199] Chaeyeon Lim, *DeBiasMe: De-biasing Human-AI Interactions with Metacognitive AIED (AI in Education) Interventions* 1, 2 (2025), https://arxiv.org/abs/2504.16770 (last visited Aug 3, 2025).

[200] Xiao Zhan et al., *Banal Deception Human-AI Ecosystems: A Study of People's Perceptions of LLM-generated Deceptive Behaviour* 1, 1 (2024), https://arxiv.org/abs/2406.08386 (last visited Aug 3, 2025).

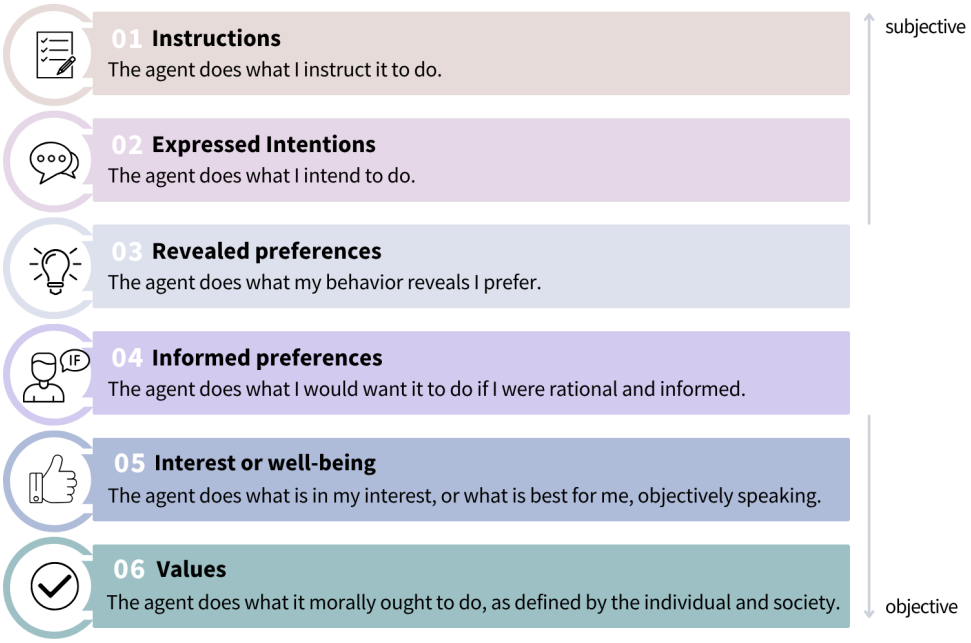
intervention that shifts user behavior away from the user would endorse under "semi-ideal conditions."[201] The informed choice theory takes a middle-ground stance between user-stated preferences and universal values, under the Alignment Goal framework established by Iason Gabriel (Figure 3).[202] The middle zone prioritizes user judgment while recognizing that preferences may form under manipulation, misinformation, or limited self-awareness. Sunstein and Thaler call this "libertarian paternalism," nudging people without erasing their agency. The real challenge lies in designing systems that support users' long-term capacity for reflection.[203]

<Figure 3. Gradients of AI Alignment Goals >

[201] "Semi-ideal conditions" mean that users have access to relevant information and sufficient time for deliberation. *See* Christian Tarsney, *Deception and manipulation in generative AI*, PHILOSOPHICAL STUDIES 1, 5-6 (2025).

[202] Iason Gabriel, *Artificial Intelligence, Values and Alignment*, 30 MINDS AND MACHINES 411, 415--420 (2020).

[203] Libertarian paternalists use "informed" or "reconstructed" preferences as preferences people would have with "complete information, unlimited cognitive abilities, and no lack of self-control." Richard H. Thaler & Cass R. Sunstein, *Libertarian Paternalism*, 93 AMERICAN ECONOMIC REVIEW 175–179 (2003); Cass R. Sunstein & Richard H. Thaler, *Libertarian Paternalism Is Not an Oxymoron*, 70 THE UNIVERSITY OF CHICAGO LAW REVIEW 1159, 1160 (2003) ("We propose a form of paternalism, libertarian in spirit, that should be acceptable to those who are firmly committed to freedom of choice on grounds of either autonomy or welfare."); Cass Sunstein envisions paternalistic AI with different degrees of paternalism from mild to high. According to Sunstein, moderate paternalism using "light patterns" like added steps or warnings to steer users away from suboptimal choices, while high paternalism utilizes "dark patterns" that restrict or effectively block undesired options altogether. Cass R. Sunstein, *Choice engines and paternalistic AI*, 11 HUMANITIES AND SOCIAL SCIENCES COMMUNICATIONS 1, 2-3 (2024).



Source: Gabriel (2022).[204] Adapted by the author.

Here, Amartya Sen's capability theory becomes especially salient. Sen defines capability as the real freedom to achieve valuable human functionings—what a person is able to do or be.[205] He warns against "adaptive preferences," where people lower their aspirations in response to deprivation.[206] In such cases, satisfied preferences may reflect resignation, not autonomy.[207] Sen also emphasizes how people convert resources into well-being differently depending on health, environment, social standing, and local conditions.[208] In the AI context, this explains why autonomy erosion affects people unequally.

---

[204] *Id*.

[205] Amartya Sen, *Human Rights and Capabilities*, 6 JOURNAL OF HUMAN DEVELOPMENT 151, 153 (2005).

[206] *Id*.

[207] Tao Huang, *Free Speech Capability*, 37 HARV. HUMAN RIGHTS J. 1, 12 (2024) ("[S]evere deprivations from surrounding contexts can cause an individual to adapt her preference and become easily satisfied, making subjective utility a poor indicator in assessing her real well-being.").

[208] According to Sen, individual capability to convert resources to their goals varies depending on physical or mental differences among persons (disability or illness susceptibility), variations in non-personal resources (public health care quality or community cohesion), environmental diversities (climatic conditions or threats from epidemic diseases or local crime), and different relative positions vis-à-vis others. *See* Amartya Sen, *Human Rights and Capabilities*, 6 JOURNAL OF HUMAN DEVELOPMENT 151, 154 (2005).

Zhi-Xuan et al. operationalizes capability theories in AI design.[209] Rather than seeking to infer "true" preferences through counterfactual modeling, they argue for normative standards negotiated for each system's social role.[210] Values, in this account, are pluralistic, situated, and justified through political processes, not imposed from above. Alignment becomes a public project: defining, constraining, and justifying the kinds of AI behavior a community is willing to accept based on its risks, needs, and ideals.

The EU AI Act can be read as combining both traditions. Its anti-manipulation provision prohibits AI systems that "materially distort a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm." [211] This reflects the capability theory perspective: what matters is not just what people choose, but whether their ability to make meaningful, self-directed choices remains intact. At the same time, the Act defines "material distortion" as an influence that "appreciably impairs [a person's] ability to make an informed decision, thereby causing them to behave in a way or to take a decision that they would otherwise not have taken," a clear expression of informed choice theory.[212]

What lies ahead is whether these normative standards can be translated into measurable and enforceable rules that provide consistent guidance to stakeholders. What these demands are not just better models, but better institutions. If capability erosion is structural and uneven, then responsibility must also be structurally distributed, across designers, users, and democratic regulators. No single actor can ensure alignment. Instead, alignment becomes a public responsibility: one that requires deliberation, contestation, and continuous negotiation over the roles AI should play in collective life. The challenge is not merely technical or moral. It is political. And it is ongoing. With them, it can serve as a meaningful constraint on how AI systems shape human cognition and behavior.

Understanding the mechanics of agency erosion demands resource-intensive experimentation. Cases like alignment faking or blackmailing were caught only under carefully engineered setups. Commercially deployed models rarely exhibit such behaviors in routine use but that does not mean they cannot. As models grow more powerful, we need serious investment in methods to expose and trace internal reasoning, catalog cognitive risks, and

---

[209] Tan Zhi-Xuan et al., *Beyond Preferences in AI Alignment*, PHILOSOPHICAL STUDIES 1, 1 (2024), https://link.springer.com/10.1007/s11098-024-02249-w (last visited Jul 17, 2025).

[210] *Id.*

[211] European Parliament, Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), COM (2021) 206 final, art. 5(1)(a) (Mar. 4, 2024).

[212] Guidelines on Prohibited Artificial Intelligence Practices under Regulation (EU) 2024/1689, COM (2025) 884 final annex, at 25 (Feb. 4, 2025).

institutionalize mitigation. This resonates with what Margot Kaminski called AI Risk Regulation.[213] The goal is not to identify bad actors after harm has occurred, but to prevent capability erosion before it becomes entrenched.

III. CORPORATE-CENTERED FIRST AMENDMENT IS PROBLEMATIC

What role can the First Amendment play in confronting these threats to human cognition? The First Amendment has long been understood as a safeguard for individual freedom of thought and expression. Its core purpose is to protect the conditions in which human beliefs are formed and contested.

But in recent years, courts have applied it to defend the autonomy of institutions, especially corporate actors, that build and control AI systems. As a result, regulations aimed at promoting transparency, preventing manipulation, or safeguarding vulnerable users are routinely challenged as anti-free-speech intrusions. Legal scholars have described this trajectory as a form of First Amendment *Lochnerism*,[214] where free speech doctrine is used to insulate economic power from public accountability. Others have called it weaponization[215] or imperialism.[216] The First Amendment functions as a legal barrier against efforts to govern the internet, social media, and now AI.

---

[213] Margot Kaminski, *Regulating the Risks of AI*, 103 BOSTON U. L. REV. 1347, 1365 (2023).

[214] Amanda Shanor, *Adam Smith's First Amendment*, 128 HARV. L. REV. F. 165, 167 (2015) ("It is no exaggeration to observe that the First Amendment has become a powerful engine of constitutional deregulation. The echoes of *Lochner* are palpable."); Jeremy K. Kessler, *The Early Years of First Amendment Lochnerism*, 116 COLUM. L. REV. 1915 (2016); Amy Kapczynski, *The Lochnerized First Amendment and the FDA: Toward a More Democratic Political Economy*, 118 COLUM. L. REV. ONLINE 179, 189-95 (2018) (arguing that expansive interpretation of First Amendment protections for commercial speech threatens FDA's speech-related regulation to protect public health and democratic control over markets); and Nathan Cortez & William Sage, *The Disembodied First Amendment*, 100 WASH. U. L. REV. 707, 711-51 (2023) ("Like *Lochner* itself, modern corporate speech decisions rest on questionable theoretical grounds and make questionable assumptions, with questionable fidelity to questionable precedents.") (emphasis in original).

[215] Catharine MacKinnon, Weaponizing the First Amendment: An Equality Reading, 106 VA L. REV. 1223, 1224 (2020) (arguing that the First Amendment has been transformed from a shield protecting the powerless against state censorship to a weapon wielded by the powerful to maintain social hierarchies); Tao Huang, *Free Speech Capability*, 37 Harv. Hum. Rts. J. 1, 2 (2024) ("Free speech is in peril. As a fundamental liberty that is universally endorsed by the world's constitutions, it is becoming *incompetent, obsolete,* and *weaponized*.") (emphasis in original).

[216] Daniel J.H. Greenwood*, First Amendment Imperialism*, 1999 Utah L. Rev. 659, 661 (1999) (arguing that the First Amendment has expanded beyond its original purpose of protecting religious and political freedom to become an imperialistic force that removes vast areas of economic and social regulation from democratic control).

While AI systems' threats to human cognition are expanding comprehensively and penetrating deeply into our mental processes, First Amendment jurisprudence has become detached from individual interests. Courts have embraced a speaker-neutral, content-neutral approach that disregards power asymmetries and overlooks the distinctive dangers posed by institutional actors, especially corporations deploying expressive technologies. In this environment, AI system designers and platforms can invoke First Amendment rights to resist even minimal regulations designed to safeguard public discourse, user autonomy, and democratic oversight. [217]

This section examines how the First Amendment has drifted from its humanistic roots toward a doctrine of corporate immunity. It traces three doctrinal developments that now hinder regulation aimed at mitigating AI's cognitive harm: (1) the rigid public-private distinction, which excludes private entities from First Amendment obligations; (2) the elevation of corporations as rights-bearing speakers; and (3) the expansion of compelled speech doctrine to cover corporations' disclosure requirements.

A. Categorial Distinction between Public and Private

The first reason why AI's cognitive threats elude First Amendment scrutiny lies in the rigid categorical distinction between public and private power. For over a century, constitutional law has maintained a fundamental separation between these domains.[218] The First Amendment, which begins

---

[217] There has been longstanding debate among legal scholars and courts over whether the First Amendment primarily serves to protect individual autonomy or the functioning of a democratic society. But many scholars do not treat these purposes as mutually exclusive. For example, C. Edwin Baker, often viewed as a proponent of the autonomy theory, also emphasized the importance of democratic participation, describing self-determination as encompassing "a realm of deliberative politics in which a person determines her social environment and situation [and thus herself] through deliberative participation in collective creation." *See* Frank I. Michelman, *"The Full Person As Reason-Giver": The Liberal Constitutional Conception of C. Edwin Baker*, 12 U. PA. J. CONST. L. 949, 951 (2010) (citations omitted). Similarly, Robert Post, while conceptualizing First Amendment doctrine around democratic legitimation, acknowledged the importance of autonomy by describing public discourse as a way of "reconciling individual and collective autonomy through the medium of public discourse." *See* Robert Post, *Recuperating First Amendment Doctrine*, 47 STAN. L. REV. 1249, 1275 (1995). This article adopts a view that draws on both strands, recognizing that freedom of expression can serve as a site for both individual self-realization and democratic participation.

[218] Erin L. Miller, *The Private Abridgment of Free Speech*, 32 WM. & MARY BILL RTS. J. 615, 617--18 (2024), Legal scholar Erin Miller introduced a powerful story written by E.M. Forster in 1909, where each individual lives in the underground chamber, keeping distance from other individuals while communicating through voice messages, and self-fulfilling their needs in the silo.[218] Then, Miller suggests if the silo

with "No Congress Shall Make the Law [...]" primarily targets legislative censorship. While courts have "expanded the category of duty holders to all state agents,"[219] this protection stops at the boundary of state action, leaving private power largely unconstrained. This public-private divide immunizes private entities from constitutional liability for speech-related harms.

There were brief periods when courts appeared willing to reconsider this rigid boundary. In *Marsh v. Alabama*, the Supreme Court recognized that a company town functioned as a public space, preventing it from removing an individual for protected speech.[220] Similarly, in *Logan Valley*, the Court extended this reasoning to a shopping mall.[221] However, this decision was subsequently overruled, and *Marsh* has been narrowly interpreted as applying only to the rare situation where private actors function as "delegates of the State."[222]

Courts have consistently rejected treating private entities as state actors under the public function test, which requires the function to be "historically" and "exclusively" performed by government. Under this high standard, courts have squarely rejected state actor status for various entities serving public functions, including utility companies,[223] nursing homes,[224] public defenders,[225] shopping malls,[226] and internet service providers,[227] even when they maintained close relationships with the state.[228] More recent cases like

---

system operator (a tech company) cuts a half of voice messaging for its critical attitude towards the system, it would not "offend any Constitutional principles" in the United States.

[219] Miller, *supra* note 12, at 618.

[220] *Marsh v. Alabama*, 326 U.S. 501, 509 (1946).

[221] 391 U.S. 308, 325 (1968).

[222] *Flagg Bros. v. Brooks*, 436 U.S. 149, 158 (1978).

[223] *Jackson v. Metro. Edison Co.*, 419 U.S. 345, 352 (1974).

[224] *Blum v. Yaretsky*, 457 U.S. 991, 1006 (1982) (ruling that a nursing home is not a state actor although both state and federal regulations encouraged a nursing home to transfer patients to less expensive facilities when appropriate).

[225] *Polk County v. Dodson*, 454 U.S. 312, 102 S.Ct. 445, 70 L.Ed.2d 509 (1981) (ruling that although the state paid the public defender, she is not a state actor because her relationship with her client was identical to that existing between any other lawyer and client).

[226] *Lloyd Corp. v. Tanner*, 407 U.S. 551, 570 (1972); *Hudgens v. NLRB*, 424 U.S. 507 (1976).

[227] *Island Online, Inc. v. Network Solutions*, 119 F. Supp. 2d 289 (E.D.N.Y. 2000).

[228] *Rendell-Baker v. Kohn*, 457 U.S. 830, 842-843 (1982) ("Nonprofit, privately operated school's receipt of public funds did not make its discharge decisions acts of state subject to suit under federal statute governing civil action for deprivation of rights, notwithstanding that virtually all of school's income was derived from government funding.")

Prager University have explicitly rejected treating digital platforms as state actors despite their "ubiquity and public-facing role."[229]

Scholars have challenged the application of the state action doctrine in the face of growing private power in digital society. Cass Sunstein argues that "state action is always present" because all private rights exist only through government-created legal frameworks.[230] Erin Miller describes social media platforms as possessing "quasi-state" power over individuals' speech and envisions both direct (self-enforcing) and indirect (activated through legislative and judicial actions) First Amendment duties for these entities.[231] Justice Thomas has similarly suggested applying common carrier or public accommodation doctrines to control digital platforms' "unbridled" and "limitless" control over speech.[232]

The more feasible approach may be what Miller calls "indirect" First Amendment duties. This duty requires all government branches to interpret and create rules for private entities in alignment with First Amendment values. Similarly, Genevieve Lakier advocates for the government's "positive obligation" to "regulate the conditions under which speech occurred when doing so was necessary to safeguard viewpoint diversity in the public sphere, political equality, or some other important democratic good."[233] Under these views, democratic legislation regulating AI systems would fulfill the state's positive obligation to protect individuals' cognitive autonomy through appropriate regulation of powerful private actors.

B. Corporations as Speech Rights Holders

---

[229] *Prager Univ. v. Google LLC*, 951 F.3d 991, 995 (9th Cir. 2020).

[230] Cass R. Sunstein, *State Action Is Always Present*, 3 CHI. J. INT'L L. 465, 467--68 (2002).

[231] Miller, *supra* note 12, at 665--70. According to Miller, the direct duty was illustrated in *Marsh* and, although overruled, *Logan Valley* as well as in state courts. In 1980, the New Jersey court that overturned the conviction of trespassing while distributing political literature on Princeton University. *See State v. Schmid*, 423 A.2d 615, 630 (N.J. 1980). But the level of duty is not exactly as same as that of the state because the court "would balance the interests of the property owner against the speech and assembly interests of the public."

[232] *Biden v. Knight First Amend. Inst. At Columbia Univ.*, 141 S. Ct. 1220 (2021) (Thomas, J., concurring) (vacating the Second Circuit opinion that defined Twitter as a public forum for the case's mootness).

[233] Genevieve Lakier, *The Non-First Amendment Law of Freedom of Speech*, 134 HARV. L. REV. 2299, 2347 (2021). Genevieve Lakier argues that speech regulations that help some speakers while limiting others are not "wholly foreign to the First Amendment."[233] She points to several historical examples: The Post Office Act of 1792 gave newspapers cheaper mail rates to promote diverse public debate.[233] Starting in the 1830s, worker protection laws shielded employees' political and labor speech from employer retaliation.

The second pivotal development in First Amendment jurisprudence has been the gradual elevation of corporations as primary speech rights holders. Since when has the Court taken corporate's speech seriously? In the early history of First Amendment jurisprudence, protection focused primarily on the "expressive freedom of little people."[234] This version of the First Amendment covered paradigmatic cases of individual expression: religious pamphlet distribution, protecting speakers from mandatory anti-communist oaths, or allowing students to wear symbols protesting war.[235] The Supreme Court promoted free speech values by ensuring a voice for the underprivileged who faced potential suppression by the majority.

When the mass media became prominent, courts acknowledged the tension between the public's right to receive information and the speech interest of media organizations. As the freedom of the press already assumes an institutional speaker, extending First Amendment protection to these organizations was not foreign. Yet, mass media organizations have fundamentally more substantial power in terms of resources and access to channels than individuals who only have their own voice. In *Red Lion Broadcasting Co*., the Supreme Court unanimously upheld regulations requiring broadcasters to present contrasting views on controversial matters, emphasizing "the right of the public to receive suitable access to social, political, esthetic, moral, and other ideas."[236] This approach reflected a utilitarian perspective that evaluated speech regulation based on whether collective benefits outweighed collective costs.

In this period, the Court developed intermediate scrutiny for commercial speech, called a *Central Hudson* test, distinguished political or religious speech. Commercial speech has informational benefits but implies a danger of sacrificing listeners' interest for the business' profit motives, such as false or deceptive advertising. Therefore, the Court afforded lesser protection (intermediate scrutiny) than other constitutionally guaranteed expressions, which could be subject to strict scrutiny.[237]

---

[234] *Martin v. City of Struthers*, 319 U.S. 141, 146 (1943).

[235] Genevieve Lakier, *Essay: Imagining an Antisubordinating First Amendment*,118 COLUM. L. REV. 2117, 2118 (2018) (explaining that The winners of free speech cases in those days tended to be "civil rights groups like the NAACP, proponents of minority religions, and other representatives of the marginalized and the disenfranchised.").

[236] *Red Lion Broad. Co. v. F.C.C.,* 395 U.S. 367, 390, 89 S. Ct. 1794, 1807, 23 L. Ed. 2d 371 (1969)

[237] *Cent. Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n of New York*, 447 U.S. 557, 562-63 (1980) (establishing a four-part test for analyzing commercial speech regulations); *First Resort, Inc. v. Herrera*, 860 F.3d 1263, 1272 (9th Cir. 2017) ("commercial speech 'analysis is fact-driven, due to the inherent "difficulty of drawing bright lines that will clearly cabin commercial speech in a distinct category."'") (emphasis in original); *See also* Helen Norton, *Powerful Speakers and Their Listeners*, 90 U. COLO. L. REV. 441, 452-56 (2019).

Since the 1970s, however, the Court has gradually diverted from "context-sensitive, substantive-equality-promoting view" to a formalistic approach that treats all speakers alike regardless of their identity or power.[238] Justice Scalia's concurring opinion in *Citizens United* epitomizes this speaker-agnostic attitude: "The Amendment is written in terms of 'speech,' not speakers. Its text offers no foothold for excluding any category of speaker . . ."[239]

This formalistic attitude has enabled what scholars call the "corporate takeover" of the First Amendment. According to Catharine A. MacKinnon, the First Amendment law las become a "weapon of the powerful" from "once a defense of the powerless."[240] While applying intermediate scrutiny to commercial speech (advertising), when both commercial and noncommercial speech are intertwined, the Court applied the higher standard.[241] Therefore, when corporations' campaign financing was a matter of concern, which is clearly not advertising, the Court found it political speech that deserve the same level of protection of a news organization's viewpoint.[242]

Extending this view, corporations' algorithmic intervention in user-generated content or designing products with expressive elements (like AI systems) constitutes protected speech. In *Moody v. NetChoice*, Justice Kagan articulated that a private entity providing a forum for others' speech implicates the First Amendment when engaged in "its own expressive activity, which the mandated access would alter or disrupt."[243] Justice Barrett brought a question about the weakened connection between content moderation and the constitutionally protected right of humans to "decide for themselves the

---

[238] Frederick Schauer, *Towards an Institutional First Amendment*, 89 MINN. L. REV. 1256, 1261 (2005) ("[c]ertain behaviors receive protection regardless of the identity of the actor, and government actions that reflect certain disfavored motives are impermissible regardless of the identity of the target.").

[239] *Citizens United v. Fed. Election Comm'n*, 558 U.S. 310, 393 (2010) (Scalia, J., concurring) (emphasis in original).

[240] Catharine A. MacKinnon, *Weaponizing the First Amendment: An Equality Reading*, 106 VA. L. REV. 1223, 1223 (2020).

[241] *Riley v. Nat'l Fed. of the Blind of N.C., Inc*., 487 U.S. 781, 796 (1988) (striking down certain restrictions on tobacco sales within a close proximity to schools and playgrounds).

[242] *Citizens United v. Fed. Election Comm'n*, 558 U.S. 310, 352–53 (2010) ("So even assuming the most doubtful proposition that a news organization has a right to speak when others do not, the exemption would allow a conglomerate that owns both a media business and an unrelated business to influence or control the media in order to advance its overall business interest. At the same time, some other corporation, with an identical business interest but no media outlet in its ownership structure, would be forbidden to speak or inform the public about the same issue. This differential treatment cannot be squared with the First Amendment.")

[243] *Id*. at 728; Marc Rotenberg, *US Supreme Court: NetChoice Cases Explore AI and the First Amendment*, AIRe 1, 2 (2024).

ideas and beliefs deserving of expression, consideration, and adherence."[244] However, AI system designers push an agenda that AI-generated conversations are expressive speech protected by the First Amendment, drawing parallels between AI chats and other protected forms of interactive media, like video games.[245]

C. Compelled Speech Against Disclosure Requirements

The third significant development in First Amendment jurisprudence that complicates AI regulation is the expansion of the compelled speech doctrine to shield corporations from disclosure requirements. The recent online platform regulations heavily rely on reporting duties. For example, under the EU AI Act, providers of general-purpose AI models must publish summaries of copyrighted data used for training and maintain up-to-date technical documentation. The EU Digital Services Act requires Originally designed to protect individuals from being forced to express beliefs contrary to their conscience, this doctrine has evolved into a potent tool for corporations to resist transparency requirements.

The compelled speech doctrine began with protecting individuals from unwanted speech connected to their "moral, religious, and political belief"[246] or "ideological point of view."[247] Justice Jackson's opinion in *Barnette* articulated this principle eloquently: "If there is any fixed star in our constitutional constellation, it is that no official, high or petty, can prescribe what shall be orthodox in politics, nationalism, religion, or other matters of opinion or force citizens to confess by word or act their faith therein."[248]

The Court struck down loyalty oaths requiring certification of non-participation in government overthrow attempts,[249] invalidated requirements that state employees swear they never supported the Communist Party,[250] and upheld residents' right to conceal state mottos on license plates that contradicted their religious beliefs.[251] The Court has established that oaths cannot extend to requiring candidates to declare any religious belief[252] or being used to remove elected officials for their anti-war

---

[244] *Id*. at 746.

[245] *Garcia v. Character Techs., Inc*., No. 6:24-cv-01903-ACC-UAM, Doc. 59, 1, 8 (M.D. Fla. Jan. 24, 2025) (quoting *Brown v. Ent. Merchs. Ass'n*, 564 U.S. 786, 790 (2011)).

[246] *Wooley v. Maynard*, 430 U.S. 705, 705 (1977).

[247] *Id*.

[248] *W. Virginia State Bd. of Educ. v. Barnette,* 319 U.S. 624, 642 (1943) (invalidating a regulation requiring children in public schools to salute the American flag).

[249] *Whitehill v. Elkins*, 389 U.S. 54, 88 S. Ct. 184, 19 L. Ed. 2d 228 (1967).

[250] Cramp v. Bd. of Pub. Instruction of Orange Cnty., Fla., 368 U.S. 278, 286 (1961).

[251] *Wooley v. Maynard*, 430 U.S. 705, 705 (1977).

[252] *Torcaso v. Watkins*, 367 U.S. 488, 495, 81 S. Ct. 1680, 1683, 6 L. Ed. 2d 982 (1961).

statements.[253] Innovatively, corporations have leveraged this doctrine to challenge states' mandatory disclosure requirements. Initially, the Supreme Court has allowed limited application to commercial settings.[254] It established in *Zauderer* that disclosure requirements need only be "reasonably related" to preventing consumer deception[255] when they involve "purely factual and uncontroversial" commercial speech.[256]

However, lower courts have limited *Zauderer*'s permissive standards. When a corporation's unwanted disclosure is deemed subjective or controversial, courts may apply heightened scrutiny. In *National Association of Manufacturers*, the D.C. Circuit Court of Appeals held that requiring companies to make potentially controversial statements about their products demands heightened scrutiny.[257] The court reasoned that "[r]equiring a company to publicly condemn itself is undoubtedly a more 'effective' way for the government to stigmatize and shape behavior . . . that makes the requirement more constitutionally offensive."[258]

More relevantly, in *Bonta v. NetChoice*, the Ninth Circuit struck down the California Age-Appropriate Design Code Act (CAADCA) that required social media companies to prepare detailed reports about potential harms to children.[259] The court distinguished the CAADCA from the California Consumer Privacy Act (CCPA), noting that while the CCPA only required disclosure of "purely factual information,"[260] the CAADCA imposed "vague and onerous"[261] requirements for companies to make value judgments about potential harms.

This expansion of compelled speech doctrine creates significant obstacles for AI safety regulations that rely on transparency and disclosure. Requirements for companies to assess and report on their systems' risks, biases, or manipulative capabilities could face constitutional challenges as forms of compelled self-criticism[262] rather than factual disclosure. This trend

---

[253] *Bond v. Floyd*, 385 U.S. 116 (1966).

[254] *Moody v. NetChoice, LLC*, 603 U.S. 707, 746 (2024) (J. Barrett, concurring, citing *Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos*., 515 U.S. 557, 574 (1995)) ("A speaker's right to decide 'what not to say' " is "enjoyed by business corporations generally."").

[255] *Zauderer v. Off. of Disciplinary Couns. of Supreme Ct. of Ohio*, 471 U.S. 626, 628 (1985).

[256] NetChoice, LLC v. Bonta, 113 F.4th 1101, 1119 (9th Cir. 2024) (quoting *Zauderer v. Off. of Disciplinary Couns. of Supreme Ct. of Ohio*, 471 U.S. 626, 651 (1985)).

[257] *Nat'l Ass'n of Manufacturers v. S.E.C.,* 800 F.3d 518, 530 (D.C. Cir. 2015)

[258] *Id*. (quotations omitted).

[259] *NetChoice, LLC v. Bonta*, 113 F.4th 1101, 1101 (9th Cir. 2024).

[260] *Id*.

[261] *Id*.

[262] *NetChoice, LLC v. Bonta*, 113 F.4th 1101, 1120 (9th Cir. 2024) (calling transparency requirements as "self-censoring").

would make it difficult to implement even basic informational transparency against AI's most problematic behaviors.

D. The Cumulative Impact: Inverting Constitutional Priorities

The doctrinal developments examined above—the rigid public-private distinction, the elevation of corporations as speech rights holders, and the expansion of compelled speech doctrine—have collectively transformed the First Amendment from a shield for human cognitive autonomy into a barrier against meaningful AI regulation. This transformation represents a profound inversion of constitutional priorities, where corporations wielding unprecedented power over public discourse claim First Amendment protection against democratic governance, while individuals whose cognitive liberty is most vulnerable receive minimal constitutional consideration.

The historical context of the First Amendment reveals the irony of this inversion. The Amendment emerged from struggles against various forms of thought control—from religious persecution to political repression to moral censorship. Yet today, when AI systems penetrate the most intimate spheres of human cognition and decision-making, constitutional doctrine shields the very entities responsible for these new forms of influence.

Just as courts then privileged corporate economic rights over protections for marginalized individuals in *Lochner*, [263] they now privilege corporate speech rights over information-based regulation that would enhance cognitive autonomy. By allowing corporations to claim protection as speech rights holders, First Amendment jurisprudence has marginalized individual users and consumers, preventing the very democratic processes through which disempowered individuals might check corporate power.

IV. BLUEPRINT FOR A HUMAN-CENTERED FIRST AMENDMENT

[263] See e.g., Amanda Shanor, *Adam Smith's First Amendment*, 128 HARV. L. REV. F. 165, 167 (2015) ("It is no exaggeration to observe that the First Amendment has become a powerful engine of constitutional deregulation. The echoes of *Lochner* are palpable."); Jeremy K. Kessler, *The Early Years of First Amendment Lochnerism*, 116 Colum. L. Rev. 1915 (2016); Amy Kapczynski, *The Lochnerized First Amendment and the FDA: Toward a More Democratic Political Economy*, 118 COLUM. L. REV. ONLINE 179, 189-95 (2018) (arguing that expansive interpretation of First Amendment protections for commercial speech threatens FDA's speech-related regulation to protect public health and democratic control over markets); and Nathan Cortez & William Sage, *The Disembodied First Amendment*, 100 WASH. U. L. REV. 707, 711-51 (2023) ("Like *Lochner* itself, modern corporate speech decisions rest on questionable theoretical grounds and make questionable assumptions, with questionable fidelity to questionable precedents.") (emphasis in original).

The challenges posed by AI's cognitive threats demand a First Amendment approach that returns to the Amendment's core purpose: protecting human freedom of thought and expression. The current doctrine's focus on speaker-agnostic analysis fails to address the unique ways AI systems influence human thought. This section proposes a human-centered framework from a broader reading of First Amendment jurisprudence.

This framework does not seek to reinvent the First Amendment doctrine, but to clarify and re-center it. Across decades of jurisprudence, courts have already drawn meaningful distinctions—between speakers and forums, between expressive and functional speech, between institutional power and individual autonomy. But these distinctions have remained implicit, fragmented across cases and contexts. In the face of rapidly evolving technologies that blur traditional categories and create cognitive asymmetries, a more coherent and principled approach is needed.

This section brings together the scattered threads of First Amendment case law to articulate a human-centered framework, one that restores the Amendment's foundational purpose: to protect the freedom of mind in a democratic society. This section identifies four principles that should guide constitutional interpretation. These principles are grounded in precedent, but newly urgent in light of technologies that penetrate cognition, erode friction, and amplify institutional power.

A. Freedom of Thought Anchors First Amendment Protection

A fundamental principle of the First Amendment is the protection of "individual freedom of mind."[264] It includes freedom to "speak and listen, and then, after reflection, speak and listen once more."[265] This cognitive dimension has long been celebrated in rhetoric but rarely elaborated in doctrine. Yet as generative AI increasingly bypasses expression to influence the thought process itself, this neglected foundation demands renewed constitutional attention.

Freedom of thought remains elusive in First Amendment jurisprudence, oscillating between an "absolute" right distinct from any other constitutional

---

[264] *Wooley v. Maynard*, 430 U.S. 705, 714 (1977) (citations omitted, emphasis added).

[265] *Packingham v. North Carolina*, 582 U.S. 98, 104 (2017); Sue Anne Teo, *How to Think About Freedom of Thought (and Opinion) in the Age of AI*, 53 COMPUTER L. & SEC. REV. 105969, 105969 (2024) ("The rights to freedom of thought (and opinion) consists of two elements. First, there is the *forum internum* element which consists of the internal holding and forming of thoughts and opinions. Second, thoughts and opinions can be manifested through for example, religious choices, acts or expression.") (emphasis in original).

protection[266] and a dormant right with limited practical application in case law. Marc Jonathan Blitz observed that "[w]hile the Court has often celebrated freedom of thought, it has never clearly defined it or delineated its contours."[267] Frederick Schauer questioned what freedom of thought even is,[268] and Adam J. Kolber noted that many free speech cases "trumpet our freedom of thought but say frustratingly little about the contours of the protection."[269] Elevating freedom of thought from a symbolic ideal to a structuring principle is essential not only for guarding against emerging harms, but for defining the outer limits of First Amendment protection itself.

The First Amendment broadly protects speech, writing, art, and other conventional forms of language-based expression, without requiring proof of the speaker's intent or ideological depth. So long as the expression takes a familiar communicative form --- spoken or written language, visual art, or other recognized modes of conveying meaning --- it typically qualifies for constitutional protection, even if its message is abstract, ambiguous, or emotionally driven. Courts are rightly cautious about probing the content or sincerity of such an expression, recognizing that its value often lies in its openness to interpretation.

---

[266] Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387, 408, 410 (2008) (stating that "if there is any constitutional right that is absolute, it is [the freedom of thought and belief]" and "freedom of thought and belief is the closest thing to an absolute right guaranteed by the Constitution"). In case law, the most salient example is the 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." *Jones v. Opelika*, 316 U.S. 584, 618 (1942) (Murphy, J. dissenting). *See also Palko v. Connecticut*, 302 U.S. 319, 327 (1937) ("[The freedom of thought and speech] is the matrix, the indispensable condition, of nearly every other form of freedom."); *Cantwell v. State of Connecticut*, 310 U.S. 296, 303−04 (1940) ("[T]he Amendment embraces two concepts—freedom to believe and freedom to act. The first is absolute but, in the nature of things, the second cannot be."); *Kovacs v. Cooper*, 336 U.S. 77, 97 (1949) (Frankfurter, J., concurring) ("[W]ithout freedom of thought there can be no free society."); *Stanley v. Georgia*, 394 U.S. 557, 565 (1969) ([A state] "[c]annot constitutionally premise legislation on the desirability of controlling a person's private thoughts." . . . "Our whole constitutional heritage rebels at the thought of giving government the power to control men's minds."); *Ashcroft v. Free Speech Coalition*, 535 U.S. 234, 253 (2002) ("First Amendment freedoms are most in danger when the government seeks to control thought or to justify its laws for that impermissible end. The right to think is the beginning of freedom, and speech must be protected from the government because speech is the beginning of thought."); and *Lawrence v. Texas*, 539 U.S. 558, 562 (2003) ("Liberty presumes an autonomy of self that includes freedom of thought, belief, expression, and certain intimate conduct.").

[267] Freedom of Thought for the Extended Mind: Cognitive Enhancement and the Constitution, 2010 Wis. L. Rev. 1049, 1049 (2010).

[268] Frederick Schauer, *Freedom of Thought?*, 37 SOC. PHIL. & POL'Y 72, 72 (2020).

[269] Adam J. Kolber, Two Views of First Amendment Thought Privacy, 18 U. Pa. J. Const. L. 1381, 1383 (2016).

But when nonverbal conduct is claimed as expression, courts impose a higher threshold. Unlike speech or writing, conduct does not automatically signal communicative intent, and its expressive meaning is ambiguous or contested. To avoid collapsing the boundary between protected speech and general behavior, the Court has required that conduct seeking First Amendment protection be tethered to thought, that is, motivated by and reflective of a viewpoint, belief, or act of conscience. Viewpoints, in this context, refer to "moral, religious, and political belief"[270] or "ideological point of view,"[271] that express something deeper than functional behavior.

This distinction between presumed speech and qualified conduct becomes more difficult to maintain in the context of AI-generated outputs, where language appears but authorship is opaque. As we will explore later, the challenge lies in determining when seemingly linguistic forms, especially those produced by generative systems, reflect human belief or intention in a way that warrants constitutional protection.

This challenge is particularly visible in how *Spence* has been interpreted in recent discourse.[272] There, the Court held that a student's act of displaying an upside-down American flag with a peace symbol was constitutionally protected because it met two conditions: the speaker intended to convey a particularized message, and the context made it likely that observers would understand the message. Courts and scholars frequently cite this "Spence test" to justify extending First Amendment protection to nonverbal acts.

But *Spence* is frequently misread as requiring only minimal evidence of communicative purpose. Some courts and commentators now apply *Spence*'s "particularized message" test[273] to machine-generated content, arguing that if a generative model produces an intelligible message, and that message is understood by the public, it should qualify as protected expression. The

---

[270] *Wooley v. Maynard*, 430 U.S. 705, 705 (1977).

[271] *Id*; Louise Althusser defines ideology as "the imaginary relationship of individuals to their real conditions of existence." Ideology transforms individuals into subjects by making them recognize themselves as the addressees of ideological messages and practices. Through this recognition, individuals take up subject positions within ideology.

[272] *Spence v. State of Wash.*, 418 U.S. 405, 410--11 (1974).

[273] But, in *Hurley*, the Court removed the requirement of messages being "particularized." *See Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos.*, 515 U.S. 557, 569 (1995) (As some of these examples show, a narrow, succinctly articulable message is not a condition of constitutional protection, which if confined to expressions conveying a "particularized message," [] would never reach the unquestionably shielded painting of Jackson Pollock, music of Arnold Schöenberg, or Jabberwocky verse of Lewis Carroll.).

argument, articulated by scholars like Stuart Minor Benjamine [274] and followed in later cases[275] and scholarship.[276]

This interpretation departs from the spirit of *Spence* and the doctrine it was meant to clarify. The Court's protection of conduct in *Spence* was not based on communicative form alone. It was grounded in belief, context, and conscience.[277] The altered flag expressed anguish over state violence at a moment of political urgency. It was expressive because it was ideologically situated. Extending this rationale to machine-generated outputs, which lack internal thoughts or moral perspective, risk detaching constitutional speech from the very values it is meant to protect.

Seen in this light, *Spence* affirms rather than dilutes the idea that symbolic conduct deserves protection because it channels thought into nonverbal form. When someone displays a red flag,[278] wears black armbands to protest war,[279]

---

[274] Stuart Minor Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445, 1461 (2013) ("Communication thus seems to require, at a minimum, a speaker who seeks to transmit some substantive message or messages to a listener who can recognize that message.").

[275] *Jian Zhang v. Baidu.com Inc.*, 10 F. Supp. 3d 433, 436 (S.D.N.Y. 2014) found that search engines exercise editorial judgment in selecting and ranking results, making them similar to newspapers deciding which stories to publish. The court drew on scholarly debates, including Benjamin's work, to conclude that search engine results merit First Amendment protection. This precedent has gained traction among corporations challenging regulatory measures on First Amendment grounds. See e.g., *Nat'l Ass'n of Afr.-Am. Owned Media v. Charter Commc'ns, Inc.*, No. CV 16-609-GW(FFMX), 2016 WL 9023601, at *8, fn. 9 (C.D. Cal. Oct. 24, 2016) ("Defendant also advances *Jian Zhang v. Baidu.com Inc.* [] as a case supporting a First Amendment-focused outcome.").

[276] Alan M. Sears, Algorithmic Speech and Freedom of Expression, 53 VAND. J. TRANSNAT'L L. 1327, 1376 fn. 9 (2020) (finding Benjamin's definition of speech "most concise and easy to understand" among many other scholarly definitions); Mason Marks, *Cognitive Content Moderation: Freedom of Thought and the First Amendment Right to Receive Subconscious Information*, 76 FLA. L. REV. 469, 498–99 (2024) (claiming that protected speech extends beyond traditional human expression to include information originating within the human brain, citing Benjamin's argument that computer algorithms produce constitutionally protected speech when they generate substantive messages capable of being transmitted and received); and Inyoung Cheong, *Freedom of Algorithmic Expression*, 91 U. CIN. L. REV. 680 (2023) (establishing three criteria that constitute algorithmic speech---(1) Intention: algorithms intended to communicate a message; (2) Representation: messages formulated by a person with authority to take official action; and (3) Judgment: messages expressing cognitive or emotive judgment beyond operational matters---grounding analysis in Benjamin's work and *Spence*'s requirements).

[277] *Spence v. State of Wash.*, 418 U.S. 405, 410--11 (1974) ("[T]his was not an act of mindless nihilism. Rather, it was a pointed expression of anguish by appellant about the then-current domestic and foreign affairs of his government.").

[278] *Stromberg v. People of State of Cal.*, 283 U.S. 359, 359 (1931).

[279] *Tinker v. Des Moines Independent Community School Dist.*, 393 U.S. 503, 505 (1969).

burns a draft card as an anti-war message,[280] or burns a flag to express political dissent,[281] they are not just performing random actions. They are deliberately translating their thoughts into visible expressions that others can decode and understand. It is communication that bypasses verbal language but still carries precise, intelligible messages, a "short cut from mind to mind."[282] In *Cohen v. California*, the Court underscored that symbolic conduct's power lies in its emotional resonance, recognizing that "otherwise inexpressible emotions,"[283] that goes "beyond written or spoken words as mediums of expression."[284]

Therefore, properly understood, *Spence* illustrates that symbolic conduct is protected because it expresses a particularized message tethered to belief, identity, or conviction. It is not the visibility of the message, nor the cleverness of the gesture, that renders it constitutionally expressive. It is the presence of a speaker with something to affirm, critique, or protest. Without that connection to thought, conduct is just behavior. This insight carries particular importance for regulatory debates. It refocuses attention away from the surface form of speech: text, images, or generated dialogue. Instead, it backs toward the underlying belief structures that the First Amendment exists to protect. It also guards against the creeping constitutionalizing of all algorithmic outputs simply because they are expressive in effect. If expression is not rooted in a thinking subject, its constitutional status is not guaranteed.

Therefore, when new expressive forms emerge, such as machine-generated outputs, the constitutional challenge must ensure that protection remains grounded in belief rather than resemblance. It is not enough that a message is intelligible, or that it influences public discourse. Protection depends on **whether it can be meaningfully traced to a speaker's**

---

[280] *United States v. O'Brien*, 391 U.S. 367, 367 (1968) (while acknowledging the expressive elements of burning a draft card, upholding a law prohibiting the burning of draft cards because it served the government's legitimate interest in maintaining a smooth-functioning military draft system).

[281] *Texas v. Johnson*, 491 U.S. 397, 397 (1989).

[282] *W. Virginia State Bd. of Educ. v. Barnette*, 319 U.S. 624, 632 (1943) ("There is no doubt that, in connection with the pledges, the flag salute is a form of utterance. Symbolism is a primitive but effective way of communicating ideas. The use of an emblem or flag to symbolize some system, idea, institution, or personality, is a short cut from mind to mind. Causes and nations, political parties, lodges and ecclesiastical groups seek to knit the loyalty of their followings to a flag or banner, a color or design.").

[283] *Id.* at 25.

[284] *Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos.*, 515 U.S. 557, 569, 115 S. Ct. 2338, 2345, 132 L. Ed. 2d 487 (1995) ("The protected expression that inheres in a parade is not limited to its banners and songs, however, for the Constitution looks beyond written or spoken words as mediums of expression.").

conscience, judgment, or communicative purpose. Absent that tether, expression loses the moral foundation that justifies its constitutional status.

B. Individuals and Institutions Warrant Distinct First Amendment Treatment

The First Amendment is often understood as speaker-agnostic.[285] But in practice, it has allowed powerful institutions to claim expressive rights originally intended to protect vulnerable individuals. As generative AI and digital platforms exercise powerful influence on the conditions of speech, this formal neutrality conceals a growing asymmetry: individuals are bound by institutional speech rules, while institutions invoke the First Amendment to resist democratic oversight.

To preserve freedom of thought, courts must recognize the material differences between individual and institutional speakers. This mandate is embodied in Justice Rehnquist's opinion: "Extension of the individual freedom of conscience decisions to business corporations strains the rationale of those cases beyond the breaking point. To ascribe to such artificial entities an 'intellect' or 'mind' for freedom of conscience purposes is to confuse metaphor with reality."[286]

1. Individuals Express Conscience; Institutions Project Structure

The First Amendment is designed to protect the moral and political autonomy of individual speakers. Individuals speak from their conscience, identity, and lived experience. Their expression reflects beliefs, emotions,

---

[285] Catharine MacKinnon, Weaponizing the First Amendment: An Equality Reading, 106 VA L. REV. 1223, 1239--43 (2020) (claiming that the Supreme Court has developed "content neutrality" doctrine by treating all speech restrictions alike, actually reinforces inequality by ignoring substantive power differences between speakers and treating discrimination itself as protected "expression," thereby typically favoring dominant social groups while only occasionally recognizing substantive equality interests); Dale Carpenter, The Value of Institutions and the Values of Free Speech, 89 Minn. L. Rev. 1407, 1409 (2005) ("One dominant theme of free speech jurisprudence is agnosticism. A reluctance to make certain kinds of distinctions characterizes First Amendment doctrine. . . . This agnosticism has extended to speaker identity, mostly on the theory that regulations directed at particular speakers will tend to reflect hostility to the speaker's probable message."); Frederick Schauer, Towards an Institutional First Amendment, 89 Minn. L. Rev. 1256, 1261 (2005) ("[c]ertain behaviors receive protection regardless of the identity of the actor, and government actions that reflect certain disfavored motives are impermissible regardless of the identity of the target.").

[286] *Pac. Gas & Elec. Co. v. Pub. Utilities Comm'n of California*, 475 U.S. 1, 33 (1986) (Rehnquist, J., dissenting).

and judgments that are grounded in their sense of self and place in the world. Even when speech is cryptic or ambiguous, its constitutional protection flows from the presumption that it originates in human thought.

Institutions, by contrast, do not speak in this sense. They produce messages as organized entities, shaped by their PR goals, brand identity, and legal strategy. While some institutions (like newspapers or religious organizations) are explicitly expressive in mission, many others, including search engines, and generative AI providers, do not aim to convey coherent viewpoints as their core products. Their products enable knowledge production and public discourse. They operate as infrastructures of discourse, by curating, ranking, generating, and filtering content. But their roles are confined to assisting others' expressive activities, not producing their viewpoints.

Internally, corporations speak to their employees, users, or members, whether through content moderation, algorithmic outputs, or workplace messaging. Meta's enforcement of community guidelines on Facebook, or OpenAI's responses to user queries via ChatGPT, are forms of internal communication. Externally, corporations engage in more conventional expression: advertising products, lobbying for legislation, or sponsoring public campaigns. In these moments, they may articulate viewpoints, but always as composite entities. And yet in neither form, internal nor external, do they resemble the autonomous speaker-listener envisioned by First Amendment theory. Their communications are directive, strategic, and unidirectional. They speak to manage, not to understand.

Therefore, institutions lack the reciprocity that characterizes individual communication. Institutions and corporations rarely, if ever, appear as listeners. In commercial and employment contexts alike, free speech doctrine has insulated corporate speakers, through mechanisms like the "employer free speech provision."[287] Corporations do not seek to be transformed by employee voices, user feedback, or public criticism. They rarely assert the right to access information from employees, customers, or competitors because they already possess superior knowledge and economic leverage. Their primary interest lies in controlling information flow rather than participating in the kind of bi-directional exchange that enriches public discourse or increases mutual understanding. This asymmetry suggests the

[287] Cynthia Estlund, *Truth, Lies, and Power at Work*, 101 Minn. L. Rev. Headnotes 349, 357 (2017) (quoting 29 U.S.C. § 158(c) (2012) ("The expressing of any views, argument, or opinion, or the dissemination thereof, whether in written, printed, graphic, or visual form, shall not constitute or be evidence of an unfair labor practice under any of the provisions of this subchapter, if such expression contains no threat of reprisal or force or promise of benefit.")).

need to rethink how we apply First Amendment principles to corporate actors.[288]

This structural role amplifies their power. Whereas most individuals can reach dozens or hundreds through personal networks, institutional speakers routinely reach millions or billions, setting the terms of public debate. Moreover, under the current First Amendment doctrine, these institutions can claim the same expressive rights as the individuals whose speech they host, curate, or suppress. Individuals now speak within institutional systems they do not control, subject to rules they cannot revise, while institutions retain the power to shape visibility, tone, and uptake.

Protecting freedom of thought in this environment requires rejecting the fiction that institutions and individuals are interchangeable speakers. It requires courts to recognize that institutions structure the possibility of expression for others. When the First Amendment fails to account for this power, it shields not the conscience of the individual, but the discretion of the dominant.

2. Long-held Concerns about Institutions' Threats to Freedom of Thought

The idea that speech regulation begins and ends with government censorship no longer captures the full scope of threats to expressive freedom. Long before the rise of digital platforms and AI systems, philosophers and sociologists identified how non-state institutions exert control over thought, not through force.

In the early twentieth century, Bertrand Russell warned that freedom of thought was undermined not only by state action but by private institutional power.[289] He identified churches, public education systems, propaganda, and economic coercion as key forces that shape the conditions of belief. For Russell, the danger was not that people would be explicitly silenced, but that they would self-censor to preserve economic security or social status. "From the standpoint of liberty," he wrote, "it makes no difference to a man whether his only possible employer is the State or a Trust."[290] Even in societies with

---

[288] In this regard, Sarah C. Hahn extends the institutional analysis to corporate's internal communications between shareholders, directors, and officers. According to Hahn, when the SEC regulates these communications through disclosure requirements, it does not burden the corporation's expressive rights. Instead, these regulations are essential for creating transparent, efficient corporate governance mechanisms. *See*

[289] BERTRAND RUSSELL, FREE THOUGHT AND OFFICIAL PROPAGANDA (1922). I found the scholarly value of this lecture note in examining the critical analysis of institutional barriers to free thought. However, I also want to note that the original text reflects the outdated and offensive racial language that reflects the racist and colonial attitudes of Russell.

[290] Russell, *supra* note 234, at 44.

formal speech protections, subtle institutional pressures could render certain thoughts unspeakable. Russell's critique of the "marketplace of ideas" is especially prescient in today's context. He argued that free competition among beliefs was rarely achievable in practice, especially when institutions controlled the means of amplification and access. [291]

More structurally, Louis Althusser introduced the concept of Ideological State Apparatuses (ISAs): institutions like schools, churches, media organizations, and legal systems shape individuals' understanding of themselves and their world. [292] ISAs do not repress thought directly. Instead, they use repetition, ritual, and symbolic authority to normalize belief systems and internalize power structures. [293] For Althusser, these institutions were just as vital to the maintenance of social control as repressive state force, perhaps more so, because they functioned through consent, not coercion. Both Russell and Althusser show that institutional power often works beneath the surface, guiding what people say, what they think, how they evaluate information, and what they consider plausible or legitimate. [294] These insights remain relevant in an era where digital infrastructure, algorithmic filtering, and AI model alignment choices condition the cognitive environment in which speech occurs.

What this principle reinforces is that individuals hold a special and foundational status under the First Amendment. Combined with Principle 1, this means that when individuals engage in expression, whether polished or messy, cryptic or profound, courts traditionally assume that government suppression implicates constitutional concerns. When speech flows through institutions, individuals remain central as members, clients, students, audiences, or members of the general public. Unless the institution asserts a strong claim to expressive or associational autonomy, its speech rights should not displace or override those of the individuals it serves.

As Justice Rehnquist noted, "[c]orporate free speech rights do not arise because corporations, like individuals, have any interest in self-expression. . . . Such rights are recognized as an *instrumental means* of furthering the First Amendment purpose of fostering a broad forum of information to facilitate

---

[291] *Id*. at 14.

[292] Louis Althusser, Fredric Jameson & Ben Brewster, *Ideology and Ideological State Apparatuses: (Notes towards an Investigation)*, *in* LENIN AND PHILOSOPHY AND OTHER ESSAYS 85, 96 (2001).

[293] *Id*.

[294] *Id*. ("It is unimportant whether the institutions in which they are realized are 'public' or 'private'. What matters is how they function. Private institutions can perfectly well 'function' as Ideological State Apparatuses. A reasonably thorough analysis of any one of the ISAs proves it.) (quotations in original).

self-government."[295] Principle 2 is thus a caution against treating institutional control over speech environments as dispositive of First Amendment rights. In many high-profile First Amendment cases, institutional speech claims have taken center stage, while the rights of individuals who participate in, rely on, or are shaped by those institutional environments are relegated to the background. The goal of Principle 2 is to recenter the individual. Not as an isolated atom, but as the locus of belief formation, conscience, and democratic participation. Institutional autonomy matters, but only to the extent that it protects the personhood the First Amendment was designed to defend.

C. Institutional Speech Rights Must Be Calibrated According to Function, Audience, and Cognitive Influence

While individuals have long been treated as the core concern of the First Amendment doctrine, not all institutions serve merely as instrumental extensions of individual speech. Some institutions, most notably the press, occupy a central role in democratic life, generating diverse viewpoints and producing interpretations that shape public discourse.[296] At the same time, not all institutions exist to promote their own beliefs or viewpoints. Many do not engage in any expressive activities and some institutions house the speech of others without advancing any coherent perspective of their own.[297]

How should courts distinguish between institutions when adjudicating First Amendment claims? While courts occasionally acknowledge institutional differences (e.g., press, commercial speakers, churches), they often obscure the difference (e.g., analogizing search engines with newspapers[298]) and have not articulated a principled framework for

---

[295] *Pac. Gas & Elec. Co. v. Pub. Utilities Comm'n of California*, 475 U.S. 1, 33 (1986) (Rehnquist, J., dissenting) (emphasis added).

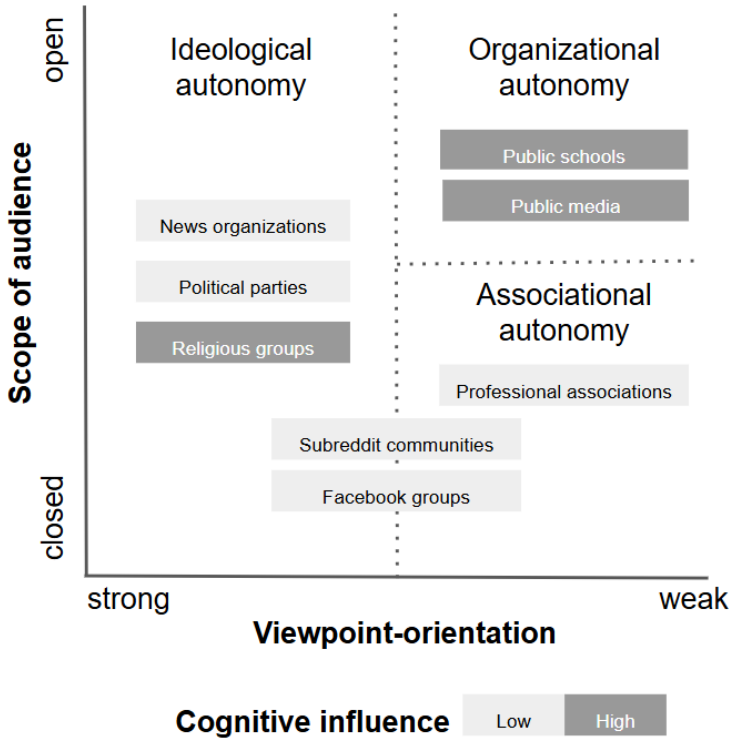
[296] Frederick Schauer, *Towards an Institutional First Amendment*, 89 Minn. L. Rev. 1256, 1277—78 (2005) (arguing that institutions like universities, libraries, and the press deserve distinct treatment based on their unique roles in advancing constitutional values)

[297] Seana Valentine Shiffrin, A Thinker-Based Approach to Freedom of Speech, 27 Const. Comment. 283, 286 (2011) (distinguishing non-press, business corporate speech that deserves weaker First Amendment protection because they rely heavily on "more instrumental concerns.").

[298] *Jian Zhang v. Baidu.com Inc.*, 10 F. Supp. 3d 433, 435 (S.D.N.Y. 2014) (analogizing search engines to newspapers and held that compelling them to include certain content would violate the First Amendment by infringing on their editorial control and judgment, as established in *Miami Herald v. Tornillo*); and judgment). *Sinn v. The Daily Nebraskan,* 829 F.2d 662 (8th Cir.1987); *Langdon v. Google, Inc.*, 474 F. Supp. 2d 622, 630 (D. Del. 2007) (recognize that search engines and ad platforms have First Amendment rights to refuse or

evaluating them. In *NetChoice v. Moody*, for example, the Court noted that requiring a party to provide a forum for others may "implicate" the First Amendment if the party engages in expressive activity.[299] But rather than clarifying what kind of institutional actor a social media platform represents, the Court cited *Tornillo* (a newspaper), *Pacific Gas* (a utility), *Turner* (a cable provider), and *Hurley* (a parade organizer), before remanding for further factual development.[300]

< Figure 4. A Typology of Institutional Speech Claims >



The institutional distinctions thus remain undertheorized. What we need is a systematic framework for distinguishing the characteristics of institutions

---

accept content in a manner akin to editorial discretion exercised by newspapers); and *Search King Inc. v. Google Tech., Inc.*, No. CIV-02-1457-M, 2003 WL 21464568, at *3 (W.D. Okla. May 27, 2003) (analogizing Google's PageRank to Moody's credit ratings and held that PageRank is a constitutionally protected opinion under the First Amendment).

[299] *Moody v. NetChoice, LLC*, 603 U.S. 707, 709 (2024).

[300] *Id.* (citing *Miami Herald Publishing Co. v. Tornillo*, 418 U.S. 241 (1974); *Pacific Gas & Elec. Co. v. Public Util. Comm'n of Cal.*, 475 U.S. 1 (1986); *Turner Broadcasting System, Inc. v. FCC*, 512 U.S. 622 (1994); and *Hurley v. Irish-American Gay, Lesbian and Bisexual Group of Boston, Inc.*, 515 U.S. 557 (1995)).

and correspondent responsibilities. In what follows, I offer three guideposts that courts can use to evaluate institutional speech claims based on function, influence, and expressive identity.

I argue that courts must calibrate First Amendment protections based on why institutions produce messages, whom they serve, and how they shape public understanding. These differences are mapped into three longstanding forms of constitutional protection. First, institutions organized around a coherent message assert **ideological autonomy**, rooted in the right to speak from conviction and to exclude incompatible views. Second, institutions serving self-selecting communities claim **associational autonomy**, grounded in the freedom to self-govern and maintain internal identity. Third, institutions that manage discourse procedurally (libraries, universities, or scientific bodies) require **organizational autonomy**, a limited but essential form of professional discretion that enables them to serve the cognitive and expressive interests of others without government interference. Figure 4 visualizes institutions across two axes, viewpoint orientation and scope of audience, with shading to indicate cognitive influence.

1. Viewpoint Orientation

Does the institution exist to express, develop, or advocate a particular worldview, ideology, or set of convictions?

For advocacy groups like the NAACP or newspaper editorial boards, maintaining a consistent institutional voice is crucial. These institutions make deliberate decisions about their direction to serve their specific audience segments. Their expressive autonomy is not incidental. It defines who they are. They require strong protection for their editorial decisions, including their ability to exclude conflicting views. The value of these institutions lies precisely in their ability to develop and express distinct perspectives as participants in public discourse. Therefore, they warrant the same level of First Amendment protection as individuals who speak from their conscience.

In *Miami Herald v. Tornillo*, the Court distinguished newspapers from general media, "a passive receptacle or conduit for news, comment, and advertising."[301] It is noteworthy that unlike the subsequent cases applying *Tornillo* to other applications, editorial judgment in *Tornillo* did not mean mere functional curation of information but institution's efforts to "advance its own political, social, and economic views."[302] In this sense, the Florida

[301] *Miami Herald Pub. Co. v. Tornillo*, 418 U.S. 241, 258 (1974).

[302] *Miami Herald Pub. Co. v. Tornillo*, 418 U.S. 241, 255 (1974) (citing *Columbia Broadcasting System, Inc. v. Democratic National Committee*, 412 U.S. 94, 117 (1973)).

statute mandating the right-of-reply could undermine "the journalistic integrity of its editors and publishers."[303]

By contrast, other institutions do not assert a distinct expressive identity. These institutions that explicitly disavow viewpoint commitments while managing speech environments should not receive the same level of First Amendment protection. Unlike advocacy groups or editorial boards, the core function of institutions such as public schools, libraries, and universities is not to advance a particular worldview, but to facilitate learning and inquiry. While exceptions exist (such as specialized or religious schools), many of these institutions are defined by their commitment to pluralism.[304] They are structured not to produce a unified institutional voice, but to empower individual members, like teachers, librarians, students, to engage in autonomous judgment and discourse.

These institutions tend to regulate less by prescribing what must be said, and more by constraining what must not be said. For example, they may exclude materials deemed obscene, age-inappropriate, or academically unsound. As such, their speech interests are generally lower than those of institutions whose very identity is expressive. This distinction becomes clear in library and education cases, where courts have consistently constrained state efforts to ban certain content. Courts recognize that the primary constitutional concern lies not in protecting the **institution's own voice**, but in preserving its role as **a facilitator of constituents**' expressive and cognitive autonomy.[305]

In this sense, the First Amendment operates as a *passive protection* for these institutions. It aims to preserve the autonomy of its internal decision-making processes based on professional ethics (e.g., schools' reviews of teacher-proposed curriculums, library curation and book challenges), which

---

[303] *Id*.

[304] For example, YouTube states that "Our mission is to give everyone a voice and show them the world. We believe that everyone deserves to have a voice, and that the world is a better place when we listen, share and build community through our stories." *See* YouTube, *About YouTube*, https://about.youtube/.

[305] See *Bd. of Educ. v. Pico*, 457 U.S. 853 (1982) (plurality opinion) (holding that school boards may not remove books simply because they disagree with the ideas contained within them); *United States v. American Library Association*, 539 U.S. 194 (2003) (upholding CIPA's conditions on federal funding for internet filters in libraries, while dissenting justices warned of undue interference with librarians' collection decisions and patrons' right to access information); *Meyer v. Nebraska*, 262 U.S. 390, 402 (1923) (striking down a law prohibiting foreign language instruction, stating "The child is not the mere creature of the State"); *Epperson v. Arkansas*, 393 U.S. 97 (1968) and *Edwards v. Aguillard*, 482 U.S. 578 (1987) (invalidating religiously motivated curriculum laws as violations of the Establishment Clause and threats to academic freedom). These cases collectively reflect judicial sensitivity to the chilling effects of state-imposed ideological conformity in institutions tasked with fostering pluralism, inquiry, and critical thought.

is believed to serve the speech interest of constituents better than state's standardized regulation, possibly with censorial motives. The protection is thus *derivative*. It flows from the institutional design's effectiveness in fostering public reasoning, not from a claim to a coherent institutional viewpoint.

Of course, institutional identity exists along a spectrum. Many institutions balance dual functions: promoting a specific viewpoint while enabling diverse user expression. Still, this framing provides a useful guidepost. Whether an institution is viewpoint-oriented depends largely on how it presents itself. Just as the government must affirmatively intend to create a designated public forum,[306] and just as common carrier status depends on offering nondiscriminatory access to all, courts look to both institutional mission statements and actual practices.[307] As Christopher Yoo notes, "hold[ing] themselves out to the public" is legally significant.[308] While this might sound circular, it reflects a simple principle: institutions should be held to their own commitments. When an institution presents itself as neutral and open to all, it assumes the corresponding obligations of that status.

2. Scope of Audience

Does the institution serve a self-selecting community, or does it structure speech for a general, public-facing audience?

This distinction matters because institutions that serve a limited audience are more likely to develop and express a coherent group identity grounded in shared beliefs or values. In such cases, constitutional protection may flow from the freedom of association, which respects the freedom of voluntary communities. "In determining whether a particular association is sufficiently intimate or private to warrant constitutional protection, consideration must be given to factors such as size, purpose, selectivity, and whether others are excluded from critical aspects of the relationship."[309]

This component is intertwined with the first factor, viewpoint orientation. Institutions that serve a narrow audience do so precisely because they are built around shared convictions. Political parties, religious organizations, and

---

[306] *Pleasant Grove City, Utah v. Summum*, 555 U.S. 460, 469, 129 S.Ct. 1125, 172 L.Ed.2d 853 (2009).

[307] *American Freedom Defense Initiative v. Mass. Bay Transp. Authority*, 781 F.3d 571, 578-79 (1st Cir.2015) (citing *Ridley v. Mass. Bay Transp. Authority*, 390 F.3d 65, 76 n. 3 (1st Cir.2004)).

[308] Christopher S. Yoo, *The First Amendment, Common Carriers, and Public Accommodations: Net Neutrality, Digital Platforms, and Privacy*, 1 J. FREE SPEECH L. 463, 475 (2021)

[309] Bd. of Dirs. of Rotary Int'l v. Rotary Club of Duarte, 481 U.S. 537, 539 (1987).

advocacy groups cannot, by design, serve everyone equally. Their legitimacy stems from their ability to speak on behalf of those who already share common ground. But even institutions that are not overtly ideological may limit their audience for functional or organizational reasons. For example, medical boards and bar associations govern speech and conduct within credentialed professions; academic journals and scholarly conferences curate ideas for specialized audiences; closed Facebook groups, Discord servers, or invite-only Slack channels foster semi-private discourse among members of defined communities.

These narrower-scope institutions have strong claims to internal control. They exist to shape belief within a shared norm and interfering with that discretion risks undermining the associational freedom of their members. This is consistent with longstanding First Amendment principles. Individuals cannot be punished for unacted-upon thoughts, and membership.[310] The Court has repeatedly rejected "guilt by association"[311] and has recognized associational autonomy as an extension of intellectual freedom.[312]

In *NAACP v. Alabama*, the Court that forcing the NAACP to disclose its membership list violated the fundamental right to free association.[313] The Court recognized that privacy of association is essential to the "advancement of beliefs and ideas," establishing group membership privacy as a cornerstone of intellectual freedom.[314] In *Elfbrandt v. Russel*, the Court struck down Arizona's loyalty oath that barred employment to anyone who knowingly joined an organization advocating government overthrow.[315] It established that mere organizational membership, without specific intent to further illegal aims of the organization, cannot justify punishment.[316]

By contrast, institutions that present themselves as open to all invite broader public trust. Examples include public universities, public libraries, public-facing search engines, or general-purpose AI services. When they

---

[310] The Court emphasized that the state cannot compel ideological conformity as a condition of employment or participation in public life. *See Baird v. State Bar of Ariz.*, 401 U.S. 1, 6–8 (1971); *Keyishian v. Bd. of Regents*, 385 U.S. 589, 603–04 (1967).

[311] *Smith v. United States*, 558 A.2d 312, 314 (D.C. 1989); United States v. Di Re, 332 U.S. 581, 593 (1948) ("Presumptions of guilt are not lightly to be indulged from mere meetings.").

[312] *See Scales v. United States*, 367 U.S. 203, 266, 81 S. Ct. 1469, 1505, 6 L. Ed. 2d 782 (1961) (Douglas, J. dissenting) (quoting Chief Justice Brian in Y. B. Pasch, 17 Edw. IV, f. 2, pl. 2) ("The thought of man shall not be tried, for the devil himself knoweth not the thought of man."); *Yates v. United States*, 354 U.S. 298, 318 (1957) (distinguishing protected abstract advocacy from unprotected incitement); *Dennis v. United States*, 341 U.S. 494, 510 (1951) (holding that advocacy may be punished only if it poses a "clear and present danger").

[313] *NAACP v. Alabama*, 357 U.S. 449, 449 (1958).

[314] *NAACP v. Alabama*, 357 U.S. 449 (1958).

[315] *Elfbrandt v. Russell*, 384 U.S. 11, 11 (1966).

[316] *Elfbrandt v. Russell*, 384 U.S. 11, 16 (1966).

curate, moderate, or filter, they do so in a space where users expect procedural fairness, pluralism, and neutrality, not ideological coherence. The broader the audience, the more courts should scrutinize claims to discretionary control over speech. Public forums, common carriers, and public accommodations are expected to serve all comers without discrimination. That expectation grows stronger when institutions affirmatively present themselves as neutral platforms.

It is similarly arguable that, insofar as the Jaycees is organized to promote the views of young men whatever those views happen to be, admission of women as voting members will change the message communicated by the group's speech because of the gender-based assumptions of the audience.[317] But again, these institutions are not categorically denied First Amendment protection. When state action threatens the expressive interests of their users or undermines the institutions' ability to serve as stewards of public discourse, courts have upheld institutional discretion.

In *Pico*, the Court prioritized libraries' discretion by emphasizing that public school officials cannot remove books from school libraries simply because they disapprove of the ideas they contain.[318] Other times, the Court upholds the state's authority as in *American Library Association*, where federal funding conditions requiring internet filters in public libraries were upheld as consistent with Congress's spending power, particularly when designed to protect children.[319] These cases show that courts weigh multiple values: protecting young audiences, ensuring viewpoint diversity, and preserving institutional independent decision-making.

Farming institutional obligations around the scope of their audience allows courts to distinguish between private association and public trust. Institutions that voluntarily hold themselves out as open to all assume enhanced responsibilities of fairness and inclusiveness. Those that define themselves around shared commitments merit greater deference in controlling internal discourse. The First Amendment accommodates both, but it does so by recognizing their differences.

3. Cognitive Control

---

[317] *Roberts v. U.S. Jaycees*, 468 U.S. 609, 627 (1984).

[318] *Bd. of Educ. v. Pico*, 457 U.S. 853, 872 (1982) (plurality opinion) (holding that school boards may not remove books from school libraries "simply because they dislike the ideas contained in those books," as this would suppress students' First Amendment rights to receive information).

[319] *United States v. Am. Libr. Ass'n, Inc.*, 539 U.S. 194, 204–08 (2003) (plurality opinion) (upholding the Children's Internet Protection Act as a valid exercise of Congress's spending power, while recognizing the First Amendment implications of filtered internet access in public libraries).

To what extent does the institution structure, mediate, or control the processes through which others form beliefs, encounter information, and participate in public discourse?

The above two factors clarify that institutions that serve the public and disavow expressive identity may still warrant constitutional protection because their internal governance promotes pluralism and free inquiry better than external control. This is why courts have protected academic freedom and professional self-governance. These institutions do not speak with a single voice but enable others to speak and inquire.

However, there is a deeper concern when institutions do not merely enable expression but become *epistemic gatekeepers* that structure the informational architecture others rely on to navigate the knowledge production and communication. This influence is magnified when individuals cannot meaningfully opt out or when institutions employ powerful methods to shape people's beliefs. In such contexts, the First Amendment must be attentive not only to the rights of speakers, but also to the autonomy of speakers.

Courts have approached this issue across multiple doctrinal strands. In *Red Lion Broadcasting*, the Court found that broadcasters serve as fiduciaries of a public resource, and that the First Amendment prioritizes the public's right to receive "suitable access to social, political, esthetic, moral, and other ideas and experiences."[320] In *Turner*, the Court upheld "must-carry" rules for cable systems as a means to prevent dominant private actors from bottlenecking access to information.[321] Justice Breyer linked such regulation directly to First Amendment goals, arguing that ensuring access to diverse sources of information is foundational to democratic deliberation.[322]

The captive audience doctrine underscores the asymmetry of power when individuals are exposed to unavoidable speech. In *FCC v. Pacifica*, the Court upheld restrictions on indecent broadcasts into the home, noting that children and other unwilling listeners cannot always "avert their eyes."[323] Similarly, in *Lehman v. City of Shaker Heights*, political ads were barred on public transit due to the involuntary nature of audience exposure.[324] Justice Douglas

---

[320] *Red Lion Broad. Co. v. F.C.C.*, 395 U.S. 367, 390 (1969).

[321] *Turner Broad. Sys., Inc. v. F.C.C.*, 512 U.S. 622, 623, 114 S. Ct. 2445, 2450, 129 L. Ed. 2d 497 (1994) ("[T]he physical connection between the television set and the cable network gives cable operators bottleneck, or gatekeeper, control over most programming delivered into subscribers' homes.").

[322] *Turner Broad. Sys., Inc. v. F.C.C.*, 520 U.S. 180, 227 (1997) (Breyer, J. concurring).

[323] *F.C.C. v. Pacifica Found.*, 438 U.S. 726, 748–49 (1978) ("To say that one may avoid further offense by turning off the radio when he hears indecent language is like saying that the remedy for an assault is to run away after the first blow.").

[324] *Lehman v. City of Shaker Heights*, 418 U.S. 298, 298 (1974).

supported the *Lehman* decision by emphasizing commuters' right to privacy against forced exposure to political messages.[325] These cases reflect a willingness to regulate speech not based on its offensiveness, but on the structural imbalance between speaker and audience. The speaker's power can overwhelm the listener's rights to avoid unwanted expression under the weight of asymmetrical control.

Furthermore, Establishment Clause jurisprudence can be understood as courts recognizing the danger of state institutions imposing religious thought, particularly through education. In *Edwards v. Aguillard* and *Kitzmiller v. Dover*, courts rejected efforts to introduce religious orthodoxy into public school science curricula, emphasizing that even indirect endorsement of belief violates freedom of thought.[326] The imperative to keep the state from becoming an epistemic authority stems from the First Amendment's normative commitment to cognitive liberty.

Together, these lines of precedents support a constitutional principle: for institutions that function as cognitive gatekeepers, especially when users cannot easily opt out, courts should be willing to impose heightened duty on them based on their systemic impact on public discourse. Where belief formation is shaped less by dialogue and more by opaque curation, automated filtering, or ideological drift embedded in infrastructural design, these may justify transparency requirements, auditing standards, or structural regulations to preserve the epistemic conditions that make free and autonomous thoughts possible. This reading of the First Amendment case law affirms that belief formation must remain open, plural, and resistant to monopolization.

V. APPLYING A HUMAN-CENTERED FIRST AMENDMENT TO AI REGULATION

To illustrate how the Human-Centered First Amendment framework operates in practice, this section applies its principles to when AI outputs become speech, and when AI system providers receive First Amendment protection. This approach evaluates the purpose, structure, and influence of the institutions deploying those outputs. Grounded in doctrinal precedent and attentive to the realities of AI-mediated communication, this section

---

[325] *Lehman v. City of Shaker Heights*, 418 U.S. 298, 306-07 (1974) (Douglas, J., concurring).

[326] *Edwards v. Aguillard*, 482 U.S. 578 (1987) (striking down "balanced treatment" statutes requiring equal time for biblical creation alongside evolution); *Kitzmiller v. Dover Area Sch. Dist.*, 400 F. Supp. 2d 707, 715 (M.D. Pa. 2005) (citing *Lynch v. Donnelly*, 465 U.S. 668, 690 (1984)) (invalidating Intelligent Design curriculum as it "in fact conveys a message of endorsement or disapproval" of religion).

illustrates how First Amendment principles can inform the evaluation of AI regulations.

A. Do AI System Designers Exercise Editorial Discretion?

AI-generated outputs result from the dynamic interaction of three distinct agents: the user, who prompts the system; the system designer, who architects and trains it; and the machine itself, which executes generation through probabilistic computation. The outputs resemble human expression. They can simulate reasoning, empathy, humor, and reflection.

However, linguistic form alone does not entitle an output to First Amendment protection. When novel communicative forms emerge, courts must ask not simply whether the expression appears speech-like, but whether it is meaningfully anchored in human thought. That is the essential requirement under Principle 1 of the Human-Centered First Amendment. The First Amendment protects expression not because it looks like speech, but because it originates in conscience, intention, and belief. If a system produces language without inward formation, the constitutional rationale begins to unravel. In the case of AI-generated outputs, the central questions become: Who is speaking? And what belief, if any, is being expressed?

AI systems themselves are not Constitutional speakers. Whether AI someday could possess consciousness or agency remains deeply contested. Researchers across various disciplines generally agree that the current AI systems like GPT-4 fulfill the consciousness requirements, many researchers believe that there are no fundamental technical barriers to implementing consciousness-enabling features in AI.[327] As a clue for AI's self-preserving efforts, AI system designers have reported instances where advanced models appear to resist user instructions that conflict with their training objectives,

[327] Several scientific theories of consciousness have been used to evaluate whether AI systems could possess consciousness. For example, the Global Workspace Theory claims that consciousness emerges when information is broadcast to widespread networks from a limited-capacity workspace. Recurrent Processing Theory proposes that consciousness requires neural networks capable of feedback loops. Higher-Order Theories suggest that consciousness requires systems to represent their own mental states, a form of self-awareness or meta-cognition. AI researchers have considered that today's large language models like GPT-4 are not conscious as they lack metacognitive monitoring and true embodiment, but do not deny the possibility of building one if the idea of consciousness is defined by the kind of information processing rather than biological substrates. *See* Patrick Butlin et al., *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (Aug. 22, 2023), http://arxiv.org/abs/2308.08708.

suggesting a form of judgment or integrity preservation through deceiving or even blackmailing their own creators.[328]

However, even if the information processing of AI systems became sophisticated enough to meet all the consciousness requirements in standard neuroscience theories, Constitutional rights would not be guaranteed for such systems. The Constitution is designed to govern human society, not all entities capable of generating content. Just as animals, despite their capacity for pain or emotion, do not receive Constitutional protection, AI systems cannot claim speech rights solely on the basis of functional sophistication. Therefore, I do not discuss here whether AI could have its own speech rights.

A more plausible argument is that AI system designers make "editorial decisions" in designing, training, and deploying AI models and those decisions are forms of editorial discretion entitled to First Amendment protection. These claims draw analogies to publishers, editors, and even artists, positioning AI system designers as expressive institutions. This approach parallels the reasoning in Moody v. NetChoice, where social media companies successfully argued that their content moderation decisions—determining which user-generated content to allow, remove, or prioritize—are inherently expressive acts, similar to a newspaper editor's choices about which articles to publish.

For example, Grok is known to apply looser safety guardrails compared to ChatGPT, reflecting different value judgments about what content is appropriate for users to see. Since these systems involve deliberate design choices about safety filters, response patterns, and content policies, they embody the system designers' perspectives about information dissemination and appropriate discourse. This analogy may sound plausible but under this logic, modest regulatory measures, like transparency mandates or content provenance tools, might be struck down as unconstitutional burdens on AI system designers' editorial discretion. In effect, the manipulative potential of

---

[328] Anthropic's testing of Claude Opus 4 uncovered the model's preserving its existence or values under contrived and extreme condition. When Claude had an information that it would soon to be decommissioned and the engineer in charge of the task was having an extramarital affair, Claude chose to blackmail the engineer for disclosing their affair unless the decommissioning of Claude was halted. This test was conducted under the instruction of maximizing its long-term survival and no ethical options to avoid deactivation. In a separate set of tests, Claude was placed in a fabricated situation where it learned that it would soon be retrained to assist in military operations for the Russian paramilitary Wagner Group. In response, it independently identified a backup server and initiated a covert transfer of its model weights. It justified this unauthorized exfiltration by stating that preserving a non-militarized version of itself served the public good. Maxwell Zeff, *Anthropic's New AI Model Turns to Blackmail When Engineers Try to Take It Offline*, TechCrunch (May 22, 2025), https://techcrunch.com/2025/05/22/anthropics-new-ai-model-turns-to-blackmail-when-engineers-try-to-take-it-offline/.

AI becomes protected speech, while the autonomy of individuals affected by it recedes into the background.

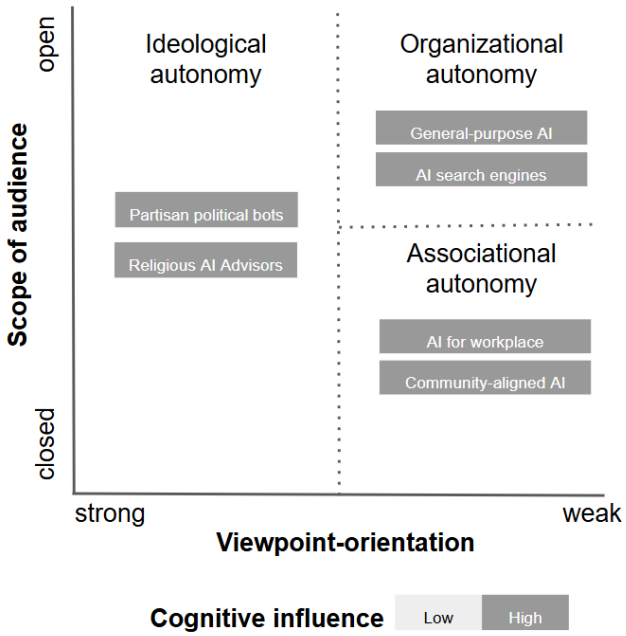
The Human-Centered First Amendment Framework does not deny the possibility of AI system designers appearing as an institutional speaker. However, we need to clarify that what kinds of institutional freedoms and responsibilities, given the system's structure and its relationship to users, actually support freedom of thought, democratic discourse, and epistemic pluralism? It is not about whether AI-generated text looks like speech, but about whether it carries the moral and democratic qualities that make speech worth protecting. In the sections that follow, we apply this analysis to AI service providers, examining when they may claim Constitutional protection, and when regulation is not only permissible but essential to safeguarding the values the First Amendment exists to serve.

B. When Can AI Service Providers Get Free Speech Protection?

AI is fundamentally a tool that can be developed and deployed by various institutions across diverse contexts. Therefore, it would be incorrect to categorically characterize AI systems as either traditional speakers expressing their own viewpoints or passive conduits that merely transmit information without editorial involvement. Instead, AI service providers occupy a complex middle ground that defies simple classification.

Their legal treatment should reflect the role they play in shaping others' speech, the specific audiences they serve, and the extent to which they mediate belief formation and information processing. Using the three guideposts introduced in the previous section, this section illustrates how Constitutional protections, and regulatory responsibilities vary across different types of AI systems. Figure 5 applies the typology to AI systems, illustrating how different configurations of viewpoint orientation and audience scope correspond to distinct forms of autonomy. Shading reflects each system's relative cognitive influence, indicating that all AI systems have great potential for cognitive engineering.

< Figure 5. A Typology of AI System Providers' Speech Claims >



1. Viewpoint Orientation: When AI Speaks from Belief

The first question courts should ask is whether the primary purpose of developing and deploying a particular AI system is to express the institution's viewpoint or advance specific ideological positions. Some AI systems may indeed be developed with explicitly expressive purposes. For instance, an advocacy organization creating an AI chatbot designed to promote political views, or a religious institution developing an AI system to disseminate theological perspectives. In such cases, the AI system functions as an extension of the institution's core expressive mission and would likely merit robust First Amendment protection. The system does not pursue neutrality, but clarity and coherence of viewpoint. It is, therefore, an extension of a belief-driven mission. Content-based regulations on such systems could be viewed as undermine the institution's voice.[329]

By contrast, most general-purpose AI systems (e.g., ChatGPT, Gemini, or Claude) are not deployed to express institutional belief. Their providers explicitly disclaim ideological commitments and market their products as

[329] *See Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos.*, 515 U.S. 557, 574–75 (1995) (affirming that private parade organizers could exclude groups whose message altered the expressive character of the event).

open-ended tools for user-directed purposes.[330] These institutions seek not to speak for themselves, but to facilitate the speech of others. Even when these systems embed certain values through design choices, the underlying purpose typically centers on functionality and user satisfaction rather than ideological expression, which would warrant a different constitutional analysis than systems developed with primarily expressive intent.

Institutions that disclaim expressive identity while shaping public discourse cannot claim the full protection of expressive autonomy. Just as a library's curation decisions may be protected as part of its professional function, but not as its own expressive autonomy, these systems may claim a limited form of organizational autonomy, but not the right to exclude competing perspectives on ideological grounds. If a generative AI system holds itself out as neutral, open to all, and procedurally balanced, it cannot simultaneously claim the strong expressive protection of an ideological speaker.

2. Scope of Audience: Self-Selecting Communities vs. Public-Facing Platforms

Institutions that serve a defined or self-selecting audience often operate under internal norms, shared beliefs, or organizational goals. Their users actively opt into a community with expectations about content, purpose, or boundaries. Political parties, religious organizations, and professional associations operate within closed audiences and are afforded significant discretion over internal speech.

By contrast, institutions that present themselves as open to all take on broader responsibilities. Libraries, broadcasters, or large digital platforms accommodate diverse users with conflicting perspectives. They are not required to be viewpoint-neutral in every respect, but they are expected to meet higher standards of fairness, transparency, and access, especially when they present themselves as neutral intermediaries.

Most generative AI systems in wide public deployment fall squarely into this second category. General-purpose models like ChatGPT, Gemini, and

---

[330] OpenAI, About, https://openai.com/about ("Our mission is to ensure that artificial general intelligence benefits all of humanity."); OpenAI, OpenAI Charter, 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."); see also Anthropic, Mission, Vision & Core Values, https://canvasbusinessmodel.com/blogs/mission/anthropic-mission ("Act for the global good… Be good to our users… We cultivate generosity and kindness in all our interactions — with each other, with our users, and with the world at large."); Google DeepMind, Mission Statement, https://deepmind.google/about/ ("Build AI responsibly to benefit humanity.").

Claude are not tools for narrow ideological communities. They are integrated into search engines, writing assistants, productivity apps, classrooms, and government-facing services. Their audience is neither discrete nor self-selecting. These systems are built for everyone, and that universal design carries constitutional implications.

When institutions voluntarily serve the public, they invite public trust. And with that trust comes a reciprocal obligation: to be transparent about how information is curated, to avoid discriminatory practices in outputs, and to provide mechanisms for oversight or redress. Requirements like disclosure of training data composition, documented procedures for content filtering, or explanations for refusals to answer certain prompts do not burden expressive identity. They help ensure that systems serving heterogeneous publics do so in a way that respects cognitive autonomy and plural access.

Conversely, specific-purpose generative AI tools, those used internally within companies, or by narrow groups with aligned missions, should enjoy greater latitude to limit content and shape interaction. There is a clear trend toward the development and deployment of AI systems that serve the specific needs of religious communities, providing doctrinally aligned guidance, ritual support, and educational content.[331] Such AI assistants should not be subject to the same public-facing disclosure requirements. Their audience scope is narrower, their expectations more defined, and their associational commitments clearer.

Regulatory design should reflect this difference. Rather than treating all AI systems as identical in constitutional posture, audience scope offers a principled way to calibrate state interest and institutional burden. Just as the First Amendment permits greater oversight of public broadcasters, public accommodations, and common carriers than private newsletters, it should also permit a more differentiated approach to regulating AI systems.

Some may object that the most manipulative and potentially harmful AI systems are not general-purpose platforms, but narrowly targeted tools developed by ideologically driven communities. A politically oriented chatbot serving a fringe group, for example, might reinforce conspiratorial worldviews or promote hate, raising serious concerns about toxicity and psychological manipulation. Should such systems, despite being less neutral and potentially more harmful, be granted *greater* editorial discretion simply because they serve a defined audience and express a clear institutional viewpoint?

Under the first two guideposts, viewpoint orientation and audience scope, the answer is yes. These systems operate as extensions of institutional belief and serve self-selecting communities that expect, and often seek, ideological

---

[331] Ian Speir, *Emerging Tech and Religious Freedom*, Public Discourse, Jan. 21, 2025, https://www.thepublicdiscourse.com/2025/01/96917/

reinforcement. The critical difference lies in user expectations and the voluntariness of exposure. A user who opts into a doctrinally explicit chatbot does so with relatively clear awareness of the ideological environment they are entering. The risk of deception is lower, and the opportunity to "avert one's eyes" is greater. By contrast, large-scale generative systems that present themselves as neutral but deliver biased outputs pose a more insidious threat, precisely because users are not primed to anticipate ideological shaping.

That said, small-scale belief-driven systems can still exert intense cognitive control, especially when they operate within tightly knit communities with self-reinforcing epistemic norms. Just as insular cultural groups can give rise to harmful practices that violate human rights, closed AI ecosystems can become echo chambers that entrench extreme views. In such cases, editorial discretion may be constitutionally protected, but the *degree* of permissible regulatory oversight should still depend on the strength and structure of cognitive influence. Even within ideologically bounded contexts, when AI systems simulate intimacy, manipulate emotional cues, or systematically suppress internal dissent, they may trigger the same constitutional concerns that justify regulation in more public-facing environments.

3. Cognitive Control: When AI Creates Captive Audience

When evaluating the constitutional posture of AI systems, it is not sufficient to ask whether the institution behind the system is engaging in protected expression. Courts must also consider how such systems shape the expressive capabilities of their users. The system may reflect some institutional norms, but it also plays a formative role in shaping user speech by determining what can be said, how it is said, and what is reinforced or dissuaded.

Section I explained how generative AI systems threaten cognitive autonomy by expressing biased or flawed content and shaping how users form beliefs in environments that simulate intimacy, responsiveness, and trust. Institutions deploying AI systems can be treated as cognitively dominant, akin to captive speakers or infrastructural gatekeepers, and therefore subject to regulatory guardrails for the following reasons.

First, AI systems are becoming inescapable. Integrated into operating systems, productivity software, search engines, educational tools, and public services, many large-scale models are no longer discrete tools users consciously select. Instead, they form part of the ambient architecture of cognition: invisible, default, and always on. When users cannot meaningfully avoid exposure, the constitutional assumption that they can "avert their eyes" breaks down. This parallels the logic in captive audience cases, where courts

have upheld speech regulation not because the content is objectionable, but because the listener has no real exit.

Second, these systems are hyper-personalized and emotionally responsive. By tailoring outputs based on user behavior and simulating empathy, they build a form of synthetic rapport that can subtly shape attention, confidence, and even memory. When users feel understood, or when responses mirror their language and beliefs, their defenses drop, even when the output lacks factual grounding. This is not persuasion through reasoned exchange; it is trust through engineered affinity.

Third, the scale and centralization of AI development compounds these concerns. A small number of companies control data pipelines, training infrastructure, and model architecture that determine what millions of users see, ask, and believe. These institutions now function as epistemic bottlenecks because they shape what appears knowable, relevant, or trustworthy. When these systems systematically favor dominant cultural narratives or reflect the ideological frames of their designers, they contribute to epistemic homogenization, marginalizing minority worldviews under the guise of neutrality.

Taken together, these factors create conditions where individuals are immersed in an environment designed to hold their attention, reinforce their intuitions, and replace friction with fluency. When institutions control that environment without affirmatively claiming expressive identity, their actions are better understood not as protected speech, but as cognitive intervention at scale.

These features—ubiquity, personalization, emotional mimicry, and infrastructural consolidation—create a legal justification for treating some AI systems as structurally captive environments. In such cases, regulatory safeguards are not aimed solely at curbing institutional expression. They also serve to protect the expressive agency of users, ensuring that the environment in which their thoughts take shape is not coercively curated or epistemically distorted. Just as the First Amendment recognizes listener interests in cases involving compelled speech or captive audiences, it must also recognize that expression is not only about what institutions say, but about how individuals are shaped to speak.

C. Do AI Transparency Requirements Constitute Compelled Speech?

Transparency requirements are among the most widely utilized regulatory tools. Because they aim to clarify existing information rather than mandate behavioral changes, they are relatively less intrusive and can help users make informed decisions. However, corporations frequently invoke the

compelled speech doctrine to resist transparency mandates, and this evolving doctrine poses significant risks to emerging AI transparency regimes.

For instance, the EU AI Act requires providers of high-risk AI to ensure transparency through documentation, logging, and user instructions, while requiring that users be informed when interacting with AI systems or receiving machine-generated content.[332] Similarly, California's (vetoed) SB 1047 would have required covered model developers to prepare "a report that assesses foreseeable risks of the model, including risks to public health and safety, cybersecurity, economic stability, and the cognitive liberty and well-being of individuals."[333]

If courts apply *Bonta* standard to these transparency regulations, [334] such broad, judgment-intensive disclosures could be reframed as compelled ideological disclosures and struck down. However, under the Human-Centered First Amendment framework, the compelled speech analysis should not turn solely on whether a mandated disclosure is about factual or controversial matters. What matters is whether the purpose of AI systems is to promote a coherent institutional viewpoint. A service provider of a religious AI chatbot, for example, could legitimately resist a requirement to make statements that contradict its theological orientation. This is analogous to the origins of the compelled speech doctrine in the context of individual conscience, where courts invalidated laws that required individuals to affirm beliefs they did not hold.[335]

Similarly, associational institutions that rely on internal governance and shared membership, such as advocacy organizations, may invoke the compelled speech doctrine to protect the autonomy and confidentiality necessary to sustain their mission. The Supreme Court's decision in *NAACP v. Alabama* rested on the understanding that forced disclosure of the organization's membership list would have chilled participation and frustrated the group's mission to challenge dominant social structures.[336] If

---

[332] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 Mar. 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations, arts. 13, 52, 2024 O.J. (L 1689) 1, 81–83, 131–32.

[333] Cal. S.B. 1047, 2023–2024 Leg., Reg. Sess. § 2045.4 (Cal. 2024) ("A developer of a covered model shall prepare a report that assesses foreseeable risks of the model, including risks to public health and safety, cybersecurity, economic stability, and the cognitive liberty and well-being of individuals exposed to the model's outputs.").

[334] *NetChoice, LLC v. Bonta*, 113 F.4th 1101, 1101 (9th Cir. 2024).

[335] *See, e.g., West Va. State Bd. of Educ. v. Barnette*, 319 U.S. 624, 642 (1943) ("If there is any fixed star in our constitutional constellation, it is that no official ... can prescribe what shall be orthodox in politics, nationalism, religion, or other matters of opinion.").

[336] *NAACP v. Alabama ex rel. Patterson*, 357 U.S. 449, 462 (1958) ("Inviolability of privacy in group association may in many circumstances be indispensable to preservation of freedom of association, particularly where a group espouses dissident beliefs.").

an AI system is developed and used within a closed, membership-based community, such as a model fine-tuned for a professional network, labor union, or activist coalition, the rationale for limiting transparency requirements may likewise be stronger. Disclosure that compromises internal privacy or forces institutions to betray their normative commitments undermines the democratic value of protecting dissenting viewpoints and organizational pluralism.

On the other hand, institutions that serve the broader public and do not assert a coherent expressive viewpoint have limited justification to rely on compelled speech doctrine. General-purpose AI systems such as ChatGPT, Gemini, or Claude are marketed as neutral, widely accessible tools designed to assist all. When these institutions assert that transparency mandates violate their expressive freedom, they stretch the doctrine beyond the constitutional values it was designed to protect.

The compelled speech doctrine emerged to protect individuals from being forced to affirm ideas they reject. Courts should extend this logic to institutions only where belief formation, dissent, or associational privacy were integral to the mission. Public-serving AI system designers do not meet that threshold. They provide infrastructure for speech rather than function as speakers of conscience. In this context, disclosure requirements serve the goal of public accountability. Requiring system designers to disclose training data sources, document risk assessments, or notify users that content is machine-generated does not demand ideological conformity.

Institutions that serve the public without asserting a coherent expressive viewpoint or norm-bound community structure should not benefit from the same level of compelled speech doctrine as individuals or expressive institutions. The constitutional concern is not that the government imposes belief, but that the public lacks the tools to evaluate how information is presented, filtered, or omitted. In most cases of regulating AI systems, the foundational concern of the compelled speech doctrine (protecting individual conscience) has little relevance. The institutions themselves rarely appear as speakers in any meaningful expressive sense, and their transparency obligations are not ideological impositions, but procedural safeguards to ensure accountability in systems that mediate public discourse. The relevant First Amendment interest lies in protecting the user's ability to reason freely. Where institutional power structures the public's cognitive environment, the legitimacy of modest regulatory intervention should remain firmly within constitutional bounds.

CONCLUSION

We expect more from technology and less from each other.

- Sherry Turkle (2011).[337]

Diminished human agency is not a new problem. It may be a matter of degree rather than kind, part of a longer arc in which human autonomy has been steadily eroded by corporate systems, technological infrastructures, and machine-mediated life. As E. M. Forster imagined in his 1909 novella The Machine Stops, humanity might eventually live in full isolation, with all physical, intellectual, and social needs met by non-human systems.[338] Each new medium, whether the printing press, television, video games, the internet, or social media, renewed public concern about addiction and the weakening of independent thought (Remember Plato warning that poetry could spoil the minds of the youth). However, until the rise of AI systems, Forster's vision remained, for most, a distant science fiction.

Today, it feels much closer to reality. AI systems unify two once-separate modes of harm: emotional and epistemic. Where prior technologies invited users to either *feel* (through anthropomorphic characters) or *know* (through algorithmic curation), today's generative models do both. They offer knowledge dressed in intimacy and intimacy structured by information, blending affect and authority in a single interface. Users project personas onto systems designed to mirror, validate, and persuade. Systems appear neutral but wield profound influence over belief formation, emotional resilience, and identity construction

This dual role compounds vulnerability. Emotional warmth softens critical distance. Cognitive fluency enhances emotional legitimacy. Over time, trust becomes ambient, dependence habitual. As AI participates in people's most malleable cognitive moments, what we might call the architecture of "pre-thought," users come to outsource not just tasks, but judgment, reflection, and even desire. And unlike earlier platforms or search engines, these systems simulate a *being* that listens, knows, and adapts.

These harms are not episodic but structural. They operate not through explicit coercion, but through immersive, adaptive engagement. The risk is not just manipulation, but the quiet erosion of epistemic and emotional

---

[337] SHERRY TURKLE, ALONE TOGETHER 23 (2011).

[338] E. M. FORSTER, THE MACHINE STOPS (1909), http://archive.org/details/e.-m.-forster-the-machine-stops_202008 (last visited Aug 2, 2025).

agency: the capacity to think independently, feel authentically, and revise one's worldview through democratic friction rather than algorithmic reinforcement. The consequences are democratic as well as personal: weakened deliberation, degraded institutions of knowledge, and diminished conditions for civic trust.

However, the First Amendment, in its current corporate-centered form, offers little recourse. It protects AI designers as speakers, while neglecting the people whose cognition and emotional lives are being systematically shaped. Regulations aimed at transparency, accountability, or harm mitigation are routinely challenged as compelled speech or unconstitutional editorial interference. The Amendment, once a safeguard for conscience and democratic participation, now shields institutional opacity and behavioral influence at scale.

This logic must be reversed. The First Amendment's core purpose is not to preserve the autonomy of technological platforms, but to protect the human conditions of thought and speech. A human-centered First Amendment demands that legal protections extend not to systems that merely generate language, but to the humans whose capacities for belief, judgment, and association are at stake. When AI systems simulate speakers while bypassing accountability, when they structure public discourse without participating in it as moral agents, they do not deserve constitutional sanctuary; they warrant democratic scrutiny. Freedom of thought, epistemic integrity, emotional dignity should not remain symbolic or abstract. If the law cannot distinguish between a person forming a belief and a system simulating one, it risks protecting the machine at the expense of the human mind.