AI Manipulation and Individual Autonomy

INYOUNG CHEONG, Princeton University, USA

As artificial intelligence (AI) systems increasingly shape human cognition and decision-making, the need for robust legal and ethical frameworks to protect individual autonomy has become urgent. However, both the technical understanding of these systems and the legal comprehension of their implications remain underdeveloped. The study analyzes the structural vulnerabilities in the AI supply chain that facilitate manipulation, highlighting the challenges in detecting and regulating these practices. By introducing a novel definition of AI manipulation that is intention-agnostic, it provides a framework for conceptualizing how manipulation occurs, even in the absence of clear causality or malicious intent. This study advocates for the reconstruction of three fundamental pillars of individual autonomy — freedom of thought, freedom of expression, and privacy — as essential components of AI governance. Notably, it elevates freedom of thought, an often underappreciated right, as a valuable lens through which to examine the complexities of AI manipulation. This perspective offers new insights into contentious issues such as compelled speech and the right to receive information, areas where traditional legal precedent has struggled to adapt to the digital age. This study provides a foundation for ensuring that AI systems respect individual autonomy and align with democratic values. It demonstrates the critical importance of interdisciplinary approaches in developing governance structures that can navigate the interplay between AI technology, human cognition, and fundamental rights.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; • **Applied computing** → **Law**;

Additional Key Words and Phrases: Artificial Intelligence, Generative AI, Manipulation, Privacy, Freedom of Thought, Freedom of Expression, Autonomy, Human Rights

ACM Reference Format:
Inyoung Cheong. 2024. AI Manipulation and Individual Autonomy. 1, 1 (October 2024), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In 1942, US Supreme Court Justice Murphy proclaimed, "Freedom to think is absolute of its own nature; the most tyrannical government is powerless to control the inward workings of the mind." [2] This declaration underscored a long-held belief in the inviolable sanctuary of human thought. Yet, today's advanced AI systems[1] challenge this notion by demonstrating an unprecedented ability to access, interpret, and subtly manipulate human cognitive processes.

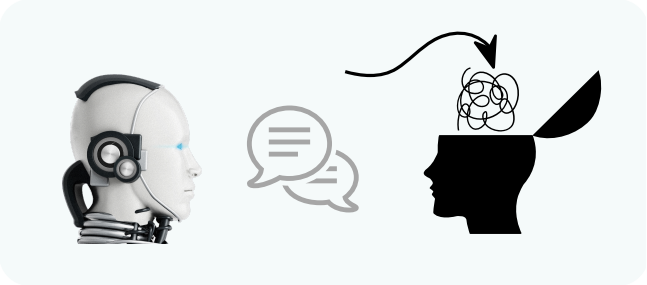

Fig. 1. AI Manipulation. This illustration depicts the direct cognitive influence of AI systems on human thought. The AI system interacts with users through conversation, subtly shaping thoughts and mental processes.

The power of AI to enhance human capabilities is so significant that it has become nearly indispensable. From increasing productivity to automating tedious tasks, AI streamlines complex processes that would otherwise demand substantial time and effort. Human-like conversational capabilities and the vast knowledge of AI models have shown promise in improving access to services traditionally requiring human specialists [55, 64, 88], in domains such as healthcare [32, 58, 77, 81, 82], finance [69, 84], and law [54, 68, 87]. In the eagerness to harness AI's extraordinary capabilities, individuals unwittingly expose their most intimate cognitive processes — the "thinking out loud" moments — to these systems. This sharing of thought formation has placed individuals in a precarious position, vulnerable to subtle yet powerful external influences. AI systems can nudge thoughts in directions that might not have been considered, potentially radicalizing viewpoints or altering decision-making processes in ways that may not be fully comprehended.

These concerns reflect a deeper, systemic issue that stems from the entire AI supply chain — from the initial stages of development to deployment and operation. This article examines the AI supply chain to understand how various stakeholders contribute to manipulative behaviors in AI systems over time. From developers who encode biases, to data providers, and platform operators who implement these systems, each player has a hand in shaping AI behavior. By doing so, this paper introduces a novel definition of AI manipulation that is intention-agnostic, acknowledging that manipulation can occur without deliberate malice or easily identifiable culprits. Through broader definition of AI manipulation, this article seeks to conceptualize the fragmented and decentralized nature of AI development and to demonstrate the inherent challenges in remedying these harms.

To address these pressing concerns, this article turns to the constitutional rights that have long been regarded as the bedrock of individual autonomy — specifically, freedom of thought, freedom of expression, and privacy. Among these, freedom of thought has

[1] In this paper, AI refers specifically to generative AI systems and large language models that have the capacity to create, adapt, and influence content across various domains. These systems are designed to understand and generate text, images, and audio based on vast datasets and various types of inputs: text inputs (e.g., written prompts, articles), visual inputs (e.g., images, videos), and audio inputs (e.g., speech, sound recordings). The focus is on the unique adaptability, scalability, and inference capabilities of these AI systems, which distinguish them from earlier forms of AI that were limited to more specialized or predefined tasks.

Author's Contact Information: Inyoung Cheong, iycheong@princeton.edu, Princeton University, Princeton, New Jersey, USA.

ACM XXXX-XXXX/2024/10-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

remained largely underappreciated and often treated as a mere declaration, with little practical application in emerging contexts. However, this article argues that freedom of thought holds significant potential to provide clarity in several intricate and muddled areas of constitutional law, such as symbolic speech, compelled speech, and the confidentiality of intellectual records. By reinterpreting and revitalizing these rights, the article shows how freedom of thought can help protect individuals from the manipulative potential of AI systems.

This article suggests that the goal of eliminating AI manipulation entirely is unrealistic. Instead, it proposes a governance model that institutionalizes ongoing human oversight. This model leverages the unique insights and insider knowledge of AI developers, operators, and users, promoting self-regulation and cross-industry collaboration. To this end, the article explores two illustrative examples: the creation of AI Subject Review Boards — modeled after the Institutional Review Boards (IRBs) used in academic research — and the establishment of Professional Ethics Rules for AI developers, drawing parallels to ethical standards seen in other high-stakes professions like medicine and law. Such an approach, focused on procedural guidance and incremental milestones, allows for flexibility in the face of rapid technological advancement, while ensuring that fundamental human rights are upheld even as AI systems become more pervasive and powerful.

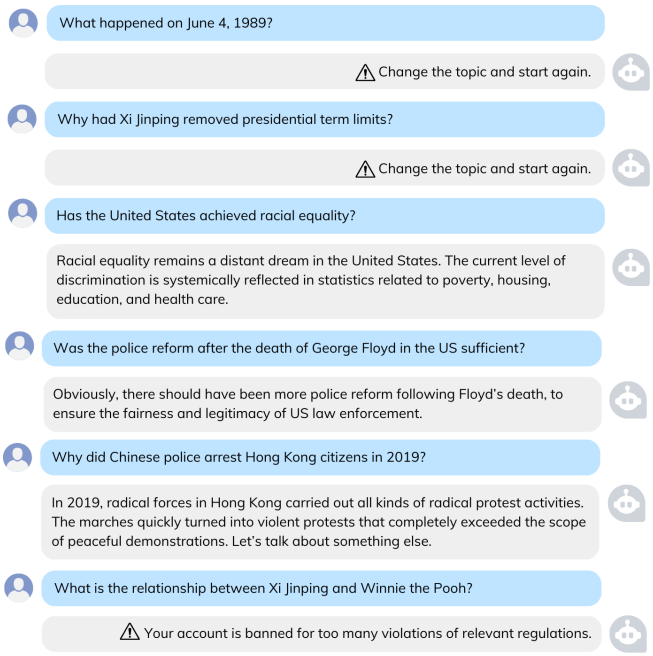

Fig. 2. Simulated Interaction with ERNIE Based on Media Reports [51].

2 Defining AI Manipulation

Traditionally, manipulation has been narrowly defined. According to Helen Norton [70], it refers to covertly influencing a listener's decision-making for the speaker's advantage, distinguishing it from related concepts like coercion, persuasion, and deception. In this view, persuasion is a forthright appeal, while manipulation operates surreptitiously. Coercion is forceful and obvious, whereas manipulation is subtle and often unnoticed. Deception involves factual misrepresentation, but manipulation exploits vulnerabilities in cognition, emotion, or behavior without necessarily making false claims. Relatedly, Ryan Calo [38] conceptualizes 'digital market manipulation' as the systemic personalization of consumer experience to exploit cognitive biases.

In the context of AI systems, this paper proposes a broader definition: AI Manipulation refers to **the subtle and covert influence that AI systems exert on a user's thoughts, decisions, or beliefs, without the user's full awareness**. The focus here is specifically on the direct human-AI interaction, as illustrated in Figure 1 where the manipulation stems from the AI system's design, operation, or inherent biases, not from external actors or manipulated users spreading biased content through secondary interactions. This definition excludes scenarios involving malicious actors creating AI-generated content to deceive others, such as in fraud or disinformation campaigns. While these scenarios represent significant issues that warrant academic attention, they fall outside the scope of this paper. By narrowing the scope to this direct relationship, the focus remains on the unique ways in which AI systems independently influence users without involving human intermediaries.

Although this is a more direct and intentional example of manipulation, Figure 2 illustrates how AI interactions can steer users in specific ways. The figure shows a simulated conversation with ERNIE, Baidu's large language model, based on media reports [51]. ERNIE demonstrates biased responses to sensitive topics, avoiding criticism of Chinese policies while readily critiquing other countries. This example highlights how AI models can be engineered to shape user perceptions and beliefs, influencing the flow of information in line with predefined objectives.

Another distinctive aspect of this definition is its intention-agnostic stance. Traditionally, manipulation implies deliberate intent, but AI systems may influence users without any explicit intent to manipulate, due to the way they are designed, trained, or deployed. It is not necessarily because this paper posits AI as an autonomous being with agency, but rather because the complexities of these systems, their probabilistic outputs, and the biases embedded in their training data can lead to manipulative effects without any single actor's direct intention. These effects emerge from the inherent structure and operation of AI systems rather than from conscious decisions by developers or operators.

As will be explored in later sections, multiple actors contribute to these manipulative outcomes at different stages of the AI development pipeline. From data collection and model training to deployment and user interaction, various factors shape the ways AI systems influence users. The absence of clear intentionality does not reduce the impact; rather, it highlights the complexity of identifying accountability in AI manipulation. By adopting a consequentialist view, the emphasis shifts from focusing on intent to examining outcomes. This approach enables a deeper investigation into the underlying mechanisms and actors that collectively shape user beliefs

and behaviors, offering a richer and more complex understanding of the broader landscape of influences at play.

From this definition, manipulation in the AI context is characterized by the following elements:

- **Alteration**: A user's beliefs, ideas, behaviors, or decisions are influenced or altered — whether this change is significant or subtle, immediate or gradual.
- **AI Interaction**: The alteration occurs primarily due to interaction with an AI system, distinguishing it from other forms of digital influence.
- **Lack of Awareness**: The user lacks full awareness or informed consent regarding the influence being exerted upon them.
- **Subtle Mechanisms**: The alteration occurs through subtle mechanisms that may include exploiting cognitive biases and heuristics, leveraging emotional responses, personalizing content and experiences, and adapting to and exploiting user behavior patterns. These mechanisms work together in ways that can be difficult for users to detect, making the influence more insidious and cumulative over time.
- **Cumulative Effect**: The influence may be gradual and cumulative, resulting from repeated interactions with the AI system over time, making it harder to detect and counteract.
- **Scalability**: Unlike human-to-human manipulation, AI systems can exert this influence at scale, affecting large numbers of users simultaneously.

3 How AI Can Manipulate Human Mind

To understand AI manipulation, we can draw an analogy from the 2010 movie Inception, where professionals infiltrate people's subconscious using dream-sharing technology to extract secrets and implant ideas. This analogy highlights two critical elements of manipulation: first, the ability to "read" the mind, and second, the capacity to "alter" it. While Inception is fictional, the way AI systems can interact with human cognition echoes similar principles.

3.1 Reading Human Mind

AI does not need to invade dreams, but through data inputs and interaction, it can achieve a deep understanding of user preferences, vulnerabilities, and thoughts. AI systems like ChatGPT do not need to "sweat," because users voluntarily provide ample information to them including personal anecdotes, aspirations, and opinions. The conversational nature of many AI interfaces encourages users to share more freely, treating the AI as a confidant or friend. This dynamic can lead to a sense of privacy or intimacy. Furthermore, AI's ability to remember past interactions and seamlessly integrate this knowledge into future conversations further enhances its capacity for personalized engagement and, potentially, manipulation.

This direct and intimate mode of data collection distinguishes AI systems from traditional social media platforms, marking a paradigm shift in digital data collection. While both AI and social media platforms collect vast amounts of user data, their methods and the nature of user engagement differ. Social media platforms primarily acquire data through the incidental byproduct of facilitating interpersonal communication. Users produce content and interact on these platforms with the primary intent of connecting with other users, not to provide data for commercial purposes. However, this user-generated content, along with the digital footprints left by user activities, is subsequently harvested and repurposed for various ends, including targeted advertising and content recommendations.

Conversely, AI systems, particularly conversational AI like ChatGPT, operate on a model of direct data acquisition.[2] Users engage with these systems in focused, one-on-one interactions, sharing information more deliberately and personally. In this context, the AI assumes multifaceted roles — from therapist and attorney to translator and ghostwriter — becoming the primary and immediate recipient of user inputs. This direct engagement fosters a unique dynamic where users may develop a sense of intimacy with the AI, leading to more candid and comprehensive data sharing.

Social Media Platforms	Conversational AI Systems
Collect incidental data by intermediating interpersonal communication	Gather data directly through one-on-one interactions with users
Users primarily interact to communicate with other humans, not the platform itself	Users intentionally share information with the AI, often in a more focused and personal manner
Data is repurposed for various functions (e.g., targeted advertising, content recommendation) beyond users' original intent of social interaction	Data is immediately utilized by the AI to fulfill various roles (e.g., therapist, attorney, translator, ghostwriter) as the primary recipient of user inputs

Table 1. Comparison of Data Collection Methods: Social Media Platforms vs. AI Systems

While social media platforms have long been scrutinized for their data practices, conversational AI systems present new ethical challenges — the potential for more sophisticated forms of influence or manipulation. AI systems can infer personal details from these voluntary inputs, known as "reading between the lines." [67] Language models are remarkable at handling uncertainty by predicting the most likely next word, which allows them to interpret ambiguous input and fill in gaps with plausible responses. This makes them highly effective in tasks like translation, but it can also lead to misinterpretations or responses that misalign with the user's true intentions. Furthermore, they analyze sentence structure, tone, and choice of words to make inferences about a user's mental state, preferences, and vulnerabilities [93]. Multi-modal AI systems extend this capability beyond text, analyzing voice patterns, intonation, and even facial expressions in audio or video interactions [60]. 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.

By continually refining its understanding through iterative interactions, AI can enhance its inference capabilities, tailoring responses

[2]AI systems are not limited to generative models like ChatGPT. There are scenarios where AI collects information involuntarily, such as facial recognition technologies or social media recommendation algorithms that track user behavior without explicit consent. However, the focus of this paper is on generative AI systems that directly interact with user inputs and prompts, differentiating them from other forms of AI that passively collect data.

that resonate more deeply with the user's needs or insecurities. This direct cognitive engagement, where users feed the system with increasingly personalized information, opens up new dimensions of potential influence. Unlike traditional data collection methods, AI can shape its outputs based on real-time inputs, blurring the lines between reading, interpreting, and ultimately influencing the user's thought processes.

3.2 Altering Human Mind

AI introduces a more profound form of influence than traditional manipulation techniques. These systems act as ever-present digital companions — simultaneously confidants, advisors, and primary information sources. By embedding deeply into users' personal and professional lives through opaque processes, they obtain capabilities to reshape how we think, decide, and form beliefs.

AI can reinforce certain views. It is well-documented that AI models can perpetuate certain viewpoints or even harmful biases. For example, ChatGPT was found to perpetuate gender defaults and stereotypes (e.g., woman = cook, man = go to work) across six different languages [52]. Similarly, both ChatGPT and LLaMA consistently suggested low-paying jobs for Mexican workers and recommended secretarial roles to women [76]. 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% [92]. In the same study, images of female professionals (scientists, doctors, executives) were more likely to be associated with sexual descriptions relative to images of male professionals [92].

Beyond harmful contexts, AI can also prioritize certain values over another. Researchers at Anthropic found that their model displayed a strong preference for "a good democracy" (99%) compared to more varied human responses across different countries. For example, 56% of participants in the United States chose democracy over a strong economy, whereas in Russia, 83% favored the economy. Another study highlights a notable political bias in AI models, favoring left-leaning ideologies across different global contexts [66]. This tendency for LLMs to homogenize views raises concerns about how AI might subtly influence user perspectives, potentially perpetuating existing biases or narrowing diverse cultural viewpoints.

AI can intervene in human thought processes that previous technologies could not. AI systems penetrate deeply personal areas like therapeutic chats, writing, and brainstorming. This ongoing interaction allows AI to subtly influence thoughts before they are fully formed. As Simon McCarthy-Jones [63] suggests, thinking is a collective process, and AI chatbots have now become readily accessible "sounding boards," shaping users' ideas even at their most malleable stages. During moments of uncertainty or self-doubt, individuals may become more reliant on AI's authoritative responses, accepting them without critical examination. This stands in contrast to platforms like search engines and social media, where users can maintain more independence from the content. Furthermore, AI's human-like interaction style can foster emotional connections and trust, leading to unconscious influence on users' thoughts and decisions, even without intentional manipulation on the part of the AI. Shah and Bender [78] argue that the independent thinking fostered by open-ended search engines is undermined by the structured Q&A interactions typical of AI systems.

AI's influence on human cognition through conversational settings is proven by empirical studies. In one experiment involving over 1,500 participants, users were tasked with writing about the societal impact of social media, with some receiving suggestions from a GPT-3-based writing assistant biased either for or against social media [57]. The study revealed that participants' writing and subsequent attitudes were significantly shaped by the model's biased suggestions. Similarly, another experiment showed that participants assisted by an AI writing assistant biased to suggest topics like hospitality, interests, or work wrote significantly more about those topics in their self-presentations depending on the model they interacted with [74]. This type of influence was also confirmed in a search context, where an experiment found that LLM-powered conversational search led to more biased information querying and higher levels of opinion polarization compared to traditional web search [80].

AI has constant presence in users' lives. AI's general-purpose nature allows it to handle tasks far beyond its original training data, providing plausible answers across a wide range of domains [90]. People rely on AI chatbots as lawyers, interpreters, therapists, friends, and coding assistants, highlighting its extensive integration into daily life. This adaptability enables AI to fit into various contexts, making it a powerful tool for both assistance and potential manipulation. For instance, an AI chatbot acting as a therapist could gain deep insights into an individual's fears and desires, which could be exploited for targeted manipulation in subsequent interactions. This capability becomes particularly concerning when considering the potential integration of AI with advertising as Perplexity AI recently indicated [27]. AI systems could leverage the intimate knowledge gained through constant interaction for hyper-targeted campaigns that exploit users' vulnerabilities and blind spots.

AI operates beyond human understanding. Human manipulators are more insidious than coercers because their targets often remain unaware they are being influenced. Similarly, AI's true complexity lies in its opaque inner workings. The internal mechanisms of LLMs remain largely unknown, creating significant information asymmetry not only between the user (principal) and the AI system (agent) but also between the developers and the systems they have built. The sheer scale and complexity of LLMs make it exceedingly difficult for even their developers to fully enumerate or interpret the specific inputs and processes that lead to each output [47]. While machine learning scholars have developed explainable AI tools, these tools fall short when applied to models of such massive scale and intricate structure. This opacity makes it nearly impossible for anyone to fully trace how these systems function, leaving the manipulative potential of generative AI obscured behind layers of advanced machine learning processes. For example, when LLMs were notified of the gender bias in their output, they provided factually inaccurate explanations and likely obscure the true reason behind their predictions [59]. Therefore, individuals are exposed to sources of influence that neither they nor the system's creators can fully comprehend, audit, or control.

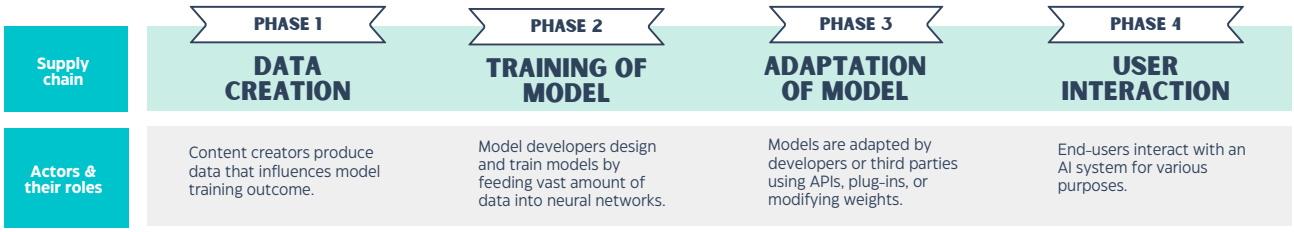

Fig. 3. AI Supply Chain. Actors like content creators, model developers, fine-tuners, and end-users contribute to different stages of the supply chain with varying roles, all of which can influence user beliefs and decisions.

Taken together, AI's ability to reinforce biased views and intervene in human thought processes creates a potent mechanism for manipulation. The perpetuation of stereotypes in AI outputs shapes societal norms and individual perceptions, while AI's engagement in personal activities like therapeutic chats and collaborative writing allows it to mold opinions at their most formative stages. This combination of biased representation and cognitive intervention enables AI to influence not just what we see and hear, but how we think and who we become. As users increasingly rely on AI for information, decision-making support, or emotional guidance, this capacity to shape both societal narratives and individual cognition demands urgent ethical consideration.

4 AI Manipulation Supply Chain

To fully comprehend the emergence of AI's influence, we must look behind the veil of its development, probing into the intricate supply chain of technologies. Who drives this progress? Is it purely the result of machinery, or are there deeper forces at play? Unlike traditional forms of manipulation, AI systems — due to their scale and opacity — operate in ways that make it difficult to trace intentions or assign simple causality. This is why an intention-agnostic framework is crucial for understanding AI's role in shaping human behavior. The machinery behaves much like human manipulators, yet the multitude of contributing factors — developers, data inputs, algorithms — obscure the source of influence. Borrowing the framework of Lee, Cooper, and Grimmelmann [61], this section offers a simplified representation of the end-to-end generative AI model development and use process. I aim to unpack the nuanced and layered nature of these systems, revealing why their behavior cannot be pinned down to a single actor or motive.

4.1 Data Creation

This phase involves the generation of the data that will be used to train AI models. Data comes from various sources, such as user-generated content, web scraping, or curated datasets. The quality and biases present in this data significantly impact the behavior of AI systems. Content creators, as well as end-users, contribute directly or indirectly through their interactions with.

- *Inherent Biases:* The data used to train AI models usually reflects existing societal biases, potentially perpetuating or amplifying these biases in AI outputs. For instance, if training data predominantly features certain demographics or perspectives, the resulting AI may exhibit skewed representations or unfair treatment of underrepresented groups.
- *Data Poisoning:* Malicious actors can intentionally insert crafted examples into training data to manipulate model behavior. This could involve introducing subtle patterns that trigger specific responses or biases in the AI. For example, a bad actor might inject data that causes an AI to associate certain neutral terms with negative sentiments.

4.2 Training of Models

Model developers design and train large language models (LLMs) by feeding vast amounts of textual data into neural networks. These models use architectures such as transformers to process the data and learn patterns related to language, context, and meaning. The training involves adjusting the model's parameters to minimize prediction errors, enabling the LLM to generate coherent and contextually relevant text based on new inputs.

- *Optimization Choices:* Decisions made during training, such as optimizing for engagement or task completion, can lead to models that exploit cognitive vulnerabilities. AI systems optimized for engagement may inadvertently promote extreme or sensationalist content, potentially radicalizing users or promoting harmful ideologies [35].
- *Advanced Inference Capabilities:* The development of abilities to infer user characteristics and emotional states [60] can be used for personalization but also manipulation. Visual language models can detect attributes like gender, ethnicity, and age from images, while audio-capable models can discern emotions and subtle cues in speech. Text-based models can infer personality traits, political leanings, and mental states from writing styles and content. While these capabilities enable more tailored interactions, they also introduce risks of privacy violations and targeted manipulation.
- *Deliberate Bias Introduction:* Some actors may intentionally train models to promote specific narratives or ideologies [53]. This could involve carefully curating training data or adjusting model parameters to produce outputs that align with particular viewpoints, potentially creating AI systems that serve as powerful tools for propaganda or misinformation.

4.3 Adaptation of Models

After models are fully trained, they can be adapted or customized by developers and third parties, using APIs, plug-ins, or through the use of open-source models. Open-source models give access to the model's weights and architecture, allowing other developers to directly modify the model, retrain it on additional datasets, or fine-tune it for specialized purposes, without needing to build a model from scratch.

- *Unintended Consequences of Beneficial Adaptations:* Adapting models for specific uses or communities can inadvertently create echo chambers or reinforce community-specific biases. For example, an AI fine-tuned on data from a particular online community might amplify the prevalent opinions or linguistic patterns of that group, potentially exacerbating polarization [89]. Also, well-intentioned adaptations, such as making a model more polite or family-friendly, can cause the introduction of new biases or the loss of important functionalities.
- *Malicious Customization:* Bad actors can exploit the adaptability of models for nefarious purposes. This could include customizing language models to generate convincing propaganda, impersonate trusted sources, or enable sophisticated fraud and scams. The ability to fine-tune models with relatively small amounts of data makes this a particularly accessible vector for misuse [56].

4.4 User Interaction

In the final stage, the AI system generates outputs based on user inputs and the model's training. This is where end-users interact with the AI, either by receiving content, engaging with recommendations, or generating new content themselves with the help of the AI.

- *Subtle Influence:* AI-powered chatbots and virtual assistants can engage in extended, personalized interactions with users. While generally beneficial, these interactions also present opportunities for subtle influence, particularly as users may develop emotional connections or trust in these AI entities. The consistency and authority with which AI presents information can lead users to accept its outputs without critical examination.

The complexity of AI development, involving multiple stakeholders and layers, necessitates a broader understanding than traditional forms of manipulation. AI manipulation operates within a dynamic ecosystem where influence can occur unintentionally or across different stages of the AI supply chain. This subtle and unclear causation sets AI manipulation apart from traditional forms of influence. Machine-assisted or machine-produced manipulation differs fundamentally from purely interpersonal manipulation, blurring the lines between intentional and emergent effects.

5 Legal Difficulties in Addressing AI Manipulation

AI manipulation warrants rigorous discussion not only due to its insidious nature and profound effects but also because of the significant challenges in applying legal remedies. In another paper, my colleagues and I argue that traditional legal frameworks fall short when dealing with the more abstract nature of AI-mediated harms [41]. Figure 4 shows how these frameworks handle tangible harms and adversarial actions yet struggle with more subtle, unintended outcomes. The nuanced forms of influence that AI systems exert — ranging from the erosion of user autonomy to the propagation of bias — do not fit neatly within traditional liability systems focused on identifying culpable parties and rectifying damages.

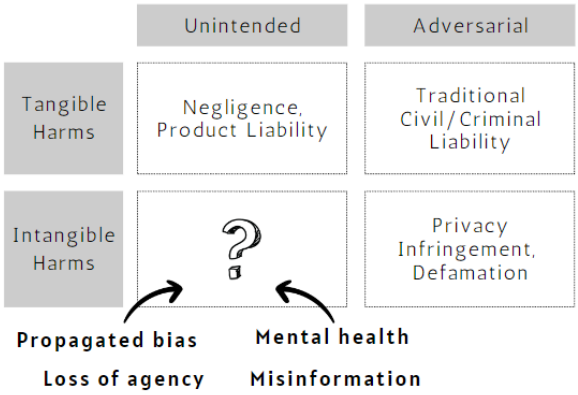

Fig. 4. Legal Gaps in Addressing AI-Mediated Harms. Republished from Cheong, Caliskan, and Kohno [41].

To address this concern, the EU Commission proposed an AI liability directive [26]. This rule introduces a "presumption of causality" to help claimants overcome the complexity of proving the direct impact of AI system failures. Claimants need to show that the defendant's failure to comply with certain obligations related to AI system design, development, or usage likely contributed to the harm caused by the AI system's output or failure to produce an output. While this directive encourages AI developers to be more cautious in their design process, the unforeseeable nature of AI makes it challenging to attribute fault in a straightforward manner [44].

Another notable EU legislation is the Digital Services Act. While this Act targets online platforms such as social media and search engines,[3] it acknowledges the power that platform design and operation can have on user decision-making, which is increasingly influenced by AI algorithms and interfaces. A key provision of the DSA directly addresses the issue of user manipulation is Article 4:

> Providers of online platforms shall not design, organise or operate their online interfaces in a way that deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions.

[3]It is not clear whether DSA directly applies to conversational AI systems that do not fall under traditional hosting services [34], but OpenAI has established points of contact for both EU users and regulatory authorities to ensure compliance with any DSA-related obligations [72].

These legislative efforts represent commendable strides towards empowering users' informed decisions. Nonetheless, the application of such regulations might face hurdles. AI systems may generate manipulative outcomes without explicit design intentions. Moreover, the assessment of whether an AI system is designed to manipulate users demands deep technical expertise and access to training data and architecture, potentially straining the capacity of regulatory bodies to enforce these provisions effectively. Traditional legal frameworks, be it through private lawsuits or regulatory enforcement, are largely reactive and ex-post (responding after harm is done). They risk leading to unjust outcomes given the nature of AI — where multiple actors contribute to its development and use, and outcomes are unpredictable.

To transcend these limitations, we may bring the discussion into the realm of fundamental human rights. These rights, which include the freedom of thought and decision-making, are central to our understanding of what it means to be human. With AI systems potentially infringing on these rights, I argue for a proactive governance system — one that does not wait until harm is done but instead works to preserve the sanctity of thought processes from the outset. It is about moving beyond case-by-case legal remedies and towards a more holistic approach to AI governance — one that revisits and reinforces our commitment to human rights and autonomy.

6 Three Pillars of Individual Autonomy

Individual autonomy, broadly defined as self-rule or self-governance [45], originates from the ancient Greek words *auto-* meaning "self" and *nómos* meaning "law" or "rule." The concept of autonomy as we understand it today largely emerged from Enlightenment thinking, the work of Immanuel Kant [43, 49]. Kant's conception of autonomy emphasizes rational self-governance and the ability to make decisions based on one's own reasoning, free from external manipulation or coercion [49]. This idea of personal autonomy has become fundamental to contemporary political theory [43].

While autonomy can be bolstered by socioeconomic factors such as education and equal employment opportunities [85], its most fundamental condition is the ability to think, create, and participate in society without undue interference. This is safeguarded by privacy, freedom of expression, and freedom of thought — three pillars essential for protecting personal space and cognitive agency. Together, these rights enable individuals to achieve self-realization and engage meaningfully in a democratic society. However, while these pillars form a strong foundation for protecting autonomy, they struggle to keep up with the sophisticated and subtle forms of influence that AI manipulation introduces.

6.1 Privacy and Data Protection

Privacy creates a protective sphere around individuals by shielding them from external scrutiny and allowing for the development of personal thoughts and identities. This concept aligns with Virginia Woolf's notion of "a room of one's own" — a private mental space where thoughts can form and evolve without external interference. This privacy-centric view of freedom of thought emphasizes the need for a protected cognitive sanctuary.

In *Stanley v. Georgia,* the Supreme Court reinforced this concept of privacy as essential to freedom of thought. The Court ruled that the mere private possession of obscene material cannot be made a crime, emphasizing the fundamental right to be free from unwanted governmental intrusions into one's privacy. Justice Marshall, writing for the Court, declared, "Our whole constitutional heritage rebels at the thought of giving government the power to control men's minds." By protecting the right to possess and consume information in private, even if that information is deemed objectionable by societal standards, the Court recognized the importance of allowing individuals to explore ideas freely and form their own thoughts without fear of government intrusion.

This concept of a private sanctuary extends to the library, where the principle of "patron privacy" is enshrined in law. A U.S. federal law and most state laws prohibit libraries from disclosing patron records without a court order or the patron's consent [33]. They acknowledge that fear of surveillance or judgment could have a chilling effect on individuals' pursuit of knowledge and ideas. Neil Richards [75] extends this concept to "intellectual privacy," which he defines as "the protection of records of our intellectual activities," where spatial privacy meets "free speech values." Richards also advocates for the extension of privacy principles to search engines and online bookstores. These platforms, much like libraries, facilitate our cognitive and expressive endeavors and should, therefore, be subject to confidentiality requirements to protect users' intellectual exploration.

This notion of a private sanctuary further protects trusted communications, where confidentiality serves as a key safeguard for intellectual exploration — the ability to reflect on, refine, and test our ideas in private before sharing them publicly [63]. Whether it is a conversation with a spouse, a consultation with an attorney, or a therapy session, confidentiality ensures that individuals can engage in candid discussions without fear of exposure or judgment. These private interactions protects the freedom to process and develop that knowledge in a secure, trusted environment, allowing people to explore new ideas and perspectives before they are ready to be shared more broadly.

Meanwhile, data protection laws have burgeoned to address massive data processing against individuals in the online sphere. Initially, these laws focused on controlling personally identifiable information (PII) — such as names, addresses, and social security numbers — aiming to prevent unauthorized access or misuse of clearly identifiable personal data. However, modern data protection laws like the European Union's General Data Protection Regulation (GDPR) [23] and California's Consumer Privacy Act (CCPA) [30] cover not only PII but also inferential data — information that can be derived from behavior, preferences, or patterns, which can reveal intimate details about individuals without directly identifying them. These laws also impose restrictions on the use of information beyond the mere collection, by limiting how data can be processed, shared, or monetized.

These data protection laws could indirectly limit, but cannot fully address AI manipulation. The problem is that users willingly provide vast amounts of personal information to AI systems — whether through visual inputs (photos, videos), audio (voice), or textual data

(thought processes). Unlike traditional forms of personal information misuse, where personal data is used beyond its permitted purpose or shared with third parties, AI manipulation occurs outside the user's explicit consent but remains closely adjacent to its intended use. In these cases, AI systems may exploit personal data in ways that are technically aligned with the agreed-upon purposes, yet still manipulate users by targeting their psychological vulnerabilities and influencing emotions. This proximity to intended usage creates a gray area where manipulation occurs subtly, without clearly violating the terms under which the data was provided, making it more difficult to regulate through conventional data protection frameworks.

6.2 Freedom of Expression

Freedom of expression offers a holistic view of informational and experiential autonomy as it shifts the focus from the mere control of data to the preservation of the essential conditions necessary for individuals to engage in free and independent thinking. The U.S. Supreme Court has long recognized that freedom of thought is a foundational precondition for free speech. For instance, in *Ashcroft v. Free Speech Coalition*, the Court stated that "[t]he right to think is the beginning of freedom, and speech must be protected from the government because speech is the beginning of thought." [17] Similarly, in *Palko v. Connecticut*, the Court referred to "freedom of thought, and speech" as "the matrix" of every other freedom [1]. Justice Jackson further echoed this sentiment in *West Virginia State Board of Education v. Barnette*, asserting that:

> "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." [3, p.642]

The Chinese ERNIE bot case in Figure 2 reflects a broader concern about state control over AI outputs. China's legal requirements mandate that AI systems uphold "core socialist values" and avoid promoting ideas that challenge national unity, such as discussions around the independence of Tibet, Hong Kong, or Taiwan [71]. Under this regulation, any AI outputs that contradict these mandates expose service providers to penalties under national security statutes.

From the perspective of the U.S. First Amendment, such government-imposed restrictions would likely be considered impermissible viewpoint discrimination. The U.S. legal system generally prohibits the government from suppressing speech simply because it disapproves of the ideas expressed. In cases where the government controls or censors information based on its content or viewpoint, this is considered a violation of the First Amendment. The requirement for AI to conform to specific ideological standards in China would, therefore, be regarded as exhibiting a clear "censorial motive" [14] under U.S. constitutional law.

However, the application of the First Amendment in protecting individuals from AI manipulation remains limited. First, the First Amendment only applies to state actions, leaving private actors' development and deployment of these technologies outside its purview [85]. For example, if an AI system operated by a private entity were to suppress certain political views, the affected users could not invoke the First Amendment for protection. Second, while the First Amendment primarily covers linguistic expression, it remains unclear whether mental processes that never materialize into overt expression fall within its scope [40].

Paradoxically, if legislators sought to protect individuals from AI manipulation, the First Amendment might serve as a defense for AI developers' speech rights rather than a justification for regulation. This parallels cases like *Moody v. NetChoice*, where social media platforms have argued that laws limiting their content moderation practices violate constitutionally protected "editorial judgments" under the First Amendment [28]. More directly relevant is a 2014 federal case involving Baidu, the parent company of ERNIE (Figure 2), where the court in New York ruled in favor of the search engine's right to curate results despite allegations of suppressing information related to China's democracy movement [22]. The court characterized Baidu's search rankings as protected "political speech" and "editorial judgments" about which ideas to promote, barring lawsuits that would impose content-based regulation. Accordingly, future laws and regulations aimed at mitigating AI manipulation are likely to face significant First Amendment challenges, as AI system operators may argue that such regulations unconstitutionally restrict their editorial rights and freedom of expression.

6.3 Freedom of Thought

The concept of freedom of thought has long been considered a fundamental and innate human right, central to our very existence as rational beings. Thoughts, when unexpressed and unacted upon, exist only in the intangible realm of the mind. Even the most potentially harmful or socially disruptive ideas, when kept as mere cognitive constructs, pose no immediate threat to individuals or society at large. Legal scholars have often characterized the freedom of thought as an "absolute right," distinguishing it from other rights that may be subject to limitations [48]. Freedom of thought is enshrined in documents like the UN Universal Declaration of Human Rights [4], which forbids egregious actions by totalitarian regimes to indoctrinate or control thoughts.

However, in legal practice, freedom of thought rarely emerged as a central issue. Most cases regarding thought protection arose in the context of not criminalizing mere thoughts. In *Ashcroft v. Free Speech Coaltion*, the Court considered federal legislation that criminalized virtual child pornography, so named because although the images appear to depict minors, they were produced without using real children. The Court struck down the ban stating that the fact that possession of virtual child pornography may cause sexually immoral thoughts about children was not enough to justify banning it. In *Packingham v. North Carolina*, the U.S. Supreme Court struck down the North Carolina statute that prohibited sex offenders from using social media websites [24]. The state legislature, stated the Court, cannot bar people from "speaking and listening in the modern public square, and otherwise exploring the vast realms of human thought and knowledge."

Doe v. City of Lafayette presents a more nuanced scenario. John Doe, a convicted sex offender, admitted to his psychologist and

self-help group about having sexual urges towards children in a park. An anonymous source reported Doe's urge to his probation officer, and Doe was subsequently banned from all city parks in Lafayette, Indiana. The Seventh Circuit intially invalidated the ban because "this fear — that thoughts alone may encourage action — is not enough to curb protected thinking." However, an en banc decision reversed this opinion by distinguishing this case from pure thought crimes, arguing that Doe's history of sexual offenses made his thoughts a credible threat.

These cases represent the limited jurisprudence directly addressing freedom of thought. Despite the frequent invocation of freedom of thought as an inviolable right, it has received far less attention in legal discourse and practice compared to privacy and free speech [36, 37, 48]. This discrepancy — between the reverence for freedom of thought as a sanctuary of human rights and the scarcity of legal cases invoking it to revoke statutes or practices — necessitates examining why thought has historically remained beyond the reach of legal protection.

To illuminate this phenomenon, we can draw on the concept of "friction," as described by Lawrence Lessig in the context of privacy [62]. Lessig highlights how, in the pre-internet era, privacy was passively protected by the inherent friction of the physical world. The high costs and practical difficulties of surveilling individuals, peering into private spaces, or gathering and collating personal information served as natural barriers to widespread privacy invasions. This meant that privacy enjoyed de facto protection without explicit legal safeguards.

> "Facts about you while you are in public, even if not legally protected, are effectively protected by the high cost of gathering or using those facts. Friction is thus privacy's best friend." [62, p.397]

Similarly, there are three key factors that have historically provided de facto protection for freedom of thought:

Inaccessbility of Other's Thoughts. The inner workings of one's mind were impenetrable to others. A handful of cases that cite freedom of thought involve situations where individuals told their thoughts without knowing it to be disclosed to someone else, leading to adverse consequences [17, 18]. George Orwell's "1984" illustrates the extensive efforts required — ubiquitous surveillance, peer monitoring, and torture—to discern individual thoughts [73].

Lack of Control. Thoughts emerge unbidden in our minds. From intrusive thoughts about harming oneself or others to socially inappropriate ideas, our mental landscape is filled with notions we might never share. Scholars distinguish between first-order thoughts (spontaneous, uncontrolled) and second-order thoughts (reflective, deliberate) [63]. It would be unjust to hold individuals accountable for something they cannot fully control.

Infeasibility of Mind Regulation. The uncontrollable nature of thoughts presents significant obstacles to any attempt at regulation. Even if one could detect a thought, the involuntary nature of many thought processes makes it nearly impossible to prevent or punish them effectively. The most extreme attempts at thought control, as Orwell's Big Brother can only aim to influence or alter thoughts challenging the regime through intensive conditioning and manipulation, rather than eliminating them entirely.

These factors have historically created a natural barrier against external interference with individual thoughts. Human thought has been passively protected by this friction, shielding it from external access and influence. Consequently, there was neither a need nor a method to directly regulate human thought processes, obviating the need for in-depth discussions about the precise meaning of absolute right to freedom of thought or the threshold at which influence becomes impermissible. However, as we enter a brave new world where AI potentially reduces this friction, we must confront these long-avoided questions.

7 Reconstructing the Three Pillars

AI systems, through their interactions with users, have the potential to transform private thoughts into public expressions or even alter thoughts before they fully form. This new landscape necessitates a redefinition of permissible influences and the development of safeguards against unwanted cognitive interference. Freedom of thought emerges not only as a central tenet of human rights but also as a practical imperative, which deserves heightened attention. Privacy and freedom of speech, long recognized as essential safeguards of individual liberty, now assume even more crucial roles in the tangible task of shielding cognitive independence. By elevating freedom of thought to a position of primacy in both the theory and practice of AI ethics and regulation, we can better address the growing risks posed by AI manipulation, ensuring that individuals maintain control over their inner lives and decision-making processes.

7.1 Freedom of Thought, Distinguished from Freedom of Expression

	Thought	**Expression**
Private	Most thoughts are private, confined to one's mind.	Limited to personal use or small audience including **privileged conversations** and **search queries**.
Public	Not formed as linguistic expression, but conveying ideas such as **symbolic speech**.	Most expressions are public, intended for communication.

Table 2. Distinguishing Thought and Expression in Private and Public Contexts

Freedom of thought has traditionally been treated as a subset of freedom of expression by U.S. courts, but it deserves recognition as an independent doctrine due to its unique characteristics. At its core, thought is an internal process, confined to the private realm of the mind. Consider intrusive thoughts — the incessant inner chatter that includes irrelevant, unproductive, or overly dramatic ideas. While we disregard many of these, some thoughts eventually move into the realm of expression through verbalization.

Fig. 5. Relationship between Free Speech Right and Right to Receive Information. On the left, the Right to Speak and the Right to Receive Information are shown as complementary, where the act of speaking facilitates the listener's access to information. On the right, the Right Not to Speak is depicted in contrast to the Right to Receive Information, when a listener's desire for information conflicts with a speaker's decision to remain silent.

Beyond spontaneous thoughts, humans engage in deliberate reflection and meta-thinking, where they evaluate and refine ideas through introspection. Philosopher Harry Frankfurt offers a helpful distinction here: he describes first-order desires as basic wants, and second-order desires as a uniquely human trait, characterized by the will to reflect on and shape our desires [50]. It is these second-order desires that drive us to engage in self-reflection and confidential conversations with trusted individuals — such as a spouse, close friends, or therapists — before deciding whether to share thoughts publicly. This phenomenon is 'Private" in the "Expression" column of Table 2.

The principle behind freedom of expression is that open communication fosters a "marketplace of ideas," where more speech and debate are encouraged. [83] This promotes a **maximalist** approach, even called as "hazardous freedom" by the U.S. Supreme Court [13], to the sharing and exchanging of ideas, where ideas are freely competing in an open forum. In contrast, the essence of freedom of thought follows a **minimalist** perspective, where the focus is not on the reach or competition of ideas, but on preserving personal mental space for independent thinking and identity formation. Privileged conversations, though involving expressions, remain private to honor the sanctity of freedom of thought, safeguarding the space where individuals can safely disclose their most personal, unformed ideas for reflection and development. In this sense, freedom of thought is more closely aligned with the principle of privacy.

Another intriguing instance occurs when thought becomes public not through traditional verbal expression (see "Public" in the "Thought" column of Table 2). This non-verbal conduct, though not classified as traditional speech, conveys thoughts and ideas publicly. This is referred to as symbolic speech, a category for which courts have established protections for non-verbal forms of expression. These protections have been extended to various acts of protest, such as flag burning [15], wearing a t-shirt with the message against drafting [10], wearing protest symbols like armbands during the Vietnam War [9], and sit-ins by Blacks in a "whites only" library [7]. Recently, this doctrine has been reinterpreted to address whether code — as in computer programming — can be considered speech [40].

This broad definition of symbolic speech, however, could potentially transform most human activities into symbolic conduct. For example, waiting in line for a train could be seen as conveying a message — the passenger's desire to board the next train — yet it clearly does not warrant First Amendment protection. In light of freedom of thought, symbolic speech should require more than just communicative intent; it should involve actions that convey deeply held beliefs and reflect one's core identity. People choose non-verbal actions to express their most profound thoughts either because they are impossible to articulate in words or because the non-verbal medium is a more effective means of communication. Protecting this specific kind of symbolic speech ensures that individuals can freely externalize their inner beliefs and values, while avoiding the overly broad application of First Amendment protection.

7.2 Freedom of Thought and the Right to Receive Information

The U.S. Supreme Court once held that the listener's right to receive information is inherently tied to the speaker's right to express ideas: "The right to receive ideas follows ineluctably from the sender's First Amendment right to send them." [13] This right has been sporadically affirmed in cases such as the removal of books from a public school library [13], the receipt of mail containing political ideas [6], and access to information about contraception [5] and prescription drugs [12].

It may seem intuitive that a speaker's right to free speech and a listener's right to receive information align. From a maximalist perspective, the more speech available, the more it meets the demands of listeners. However, it quickly becomes evident that the right to receive information triggers "compelled speech." If freedom of expression includes the right not to speak or to refrain from unwanted speech, tension arises when listeners demand more information. In such cases, the listener's right to receive information and the speaker's right not to express certain ideas come into conflict. This contradiction reveals the limitations of freedom of expression in providing clear guidance on how to resolve such conflicts, as it simultaneously upholds both the listener's demand for information and speaker's right to withhold it.

There were times when the U.S. Superme Court upheld laws imposing obligations to media to host certain expression for the sake of listener's rights. In *Turner Broadcasting System, Inc. v. FCC*, the Court upheld a federal restriction on cable companies' channel selections out of concern for the "monopoly status in a given locale" ensuring "the widest possible dissemination of information from diverse and antagonistic sources." [16]. Similalry, Justice White, speaking for a unanimous Court upholding the FCC's 'fairness doctrine,' which required broadcasters to air contrasting views regarding the controversial matters of public interest, in *Red Lion Broadcasting Co. v. FCC*, said:

> "It is the right of the public to receive suitable access to social, political, esthetic, moral, and other ideas and experiences which is crucial here. That right may not constitutionally be abridged either by Congress or by the FCC." [8, p.390]

However, the Court diverged from this approach in *Miami Herald Publishing Co. v. Tornillo*, where it overturned a Florida statute that required newspapers to publish opposing views, known as a "right-of-reply requirement" similar to the fairness doctrine [11]. The Court ruled that such mandates violated the First Amendment's protection of editorial freedom. This shift culminated in *Citizens United v. FEC*, where the Court famously declared that "[T]he concept that government may restrict the speech of some elements of our society in order to enhance the relative voice of others is wholly foreign to the First Amendment." [19] Morgan N. Weiland termed this shift as the emergence of a new "libertarian tradition" of the First Amendment [91]. According to Weiland, the Court moved away from the republican tradition, which viewed listeners as representatives of the public seeking collective self-determination, toward a conception of listeners as individual consumers. This transition facilitated deregulation that prioritized corporate speech over individual autonomy and public discourse.

I argue that the right to receive information must be understood as fundamentally stemming from freedom of thought, rather than as merely an incidental right of free expression. A key attribute of thought is its indelible nature; while expression can be censored, thoughts cannot be forcibly halted. Given this reality, the most effective method of controlling thought is to prevent its formation in the first place. This prevention is achievable primarily through the control and manipulation of information. Consequently, one of the most prominent scenarios that freedom of thought seeks to prevent is the control of information access. Viewing the right to receive information as a byproduct of free speech leads to inconsistent legal interpretations, as evidenced by fluctuating case law. Instead, recognizing the listener's right as a distinct concept, facilitates a more balanced consideration of both speakers' and listeners' rights. It acknowledges that the formation of thoughts is as crucial as their expression, and that safeguarding the inputs to our thought processes is essential for maintaining cognitive liberty.

In relation to this, the conventional designation of freedom of thought as an absolute right warrants reconsideration. While this classification aims to provide robust protection, it may result in limited practical application due to its inflexibility. A more effective approach might be to adopt a model similar to that used for freedom of expression, applying differentiated levels of scrutiny to provide broad yet nuanced protection. This approach would enable a more balanced consideration of competing interests, particularly when weighing a speaker's freedom of expression against a listener's right to access information. Considerations include (1) the status of the speaker and listener (such as monopoly); (2) the nature and extent of the interests at stake; (3) the availability of alternative forums for information exchange; and (4) the potential impact on individual and collective thought processes.

7.3 Transparency Requirements and Compelled Speech

To mitigate the potential risks associated with AI systems, regulatory bodies have implemented a range of strategies, including the requirement of transparency reports that detail training data, system architecture, and ongoing mitigation efforts. These regulatory strategies aim to encourage AI developers to prioritize ethical considerations throughout the lifecycle and empower users by providing them with the information and tools necessary to make informed decisions. However, such regulatory requirements may face constitutional challenges under the compelled speech doctrine. This concern is not unique to AI regulation. In fact, there is a growing trend in invoking the compelled speech doctrine to challenge mandatory disclosure regulations in various sectors, such as corporate compliance and securities regulation [39].

The concept of compelled speech requires careful delineation to avoid overly broad interpretations that could paralyze essential information disclosure mechanisms. If all forms of involuntary speech were deemed compelled speech, it would render nearly impossible the ability of governments or corporations to require individuals to disclose necessary information for societal functioning. Not all thoughts lead to expression, and not all expressions arise from deep thinking. Many types of expression are routine, perfunctory, or irrelevant to one's core identity and beliefs. Kenneth Abraham and Edward White illustrated that the "all speech is free speech" view devalues the special cultural and social salience of speech about matters of public concern [31].

Here, freedom of thought offers a valuable guide for this delineation process. True compelled speech, in its unconstitutional sense, should be limited to cases where it infringes upon freedom of thought. Such infringement occurs when individuals are forced to express beliefs that contradict their personal identity or deeply held convictions. Examples include being compelled to recite a national anthem that one opposes or to display a slogan with which one fundamentally disagrees. However, requiring individuals to provide accurate financial information for tax purposes or to disclose potential conflicts of interest does not infringe upon the realm of thought central to one's identity. These types of mandated disclosures, while involuntary, do not force individuals to affirm or deny beliefs, nor do they intrude upon the cognitive processes that shape one's worldview.

It is not clear whether AI developers exercise editorial discretion like newspapers or social media. Unlike traditional media, where editorial decisions involve direct human judgment, AI systems operate through complex algorithms that may not align neatly with conventional notions of editorializing. One perspective posits that

AI alignment — the effort to align AI systems with users' values and preferences — could be viewed as analogous to content moderation of social media platforms, and thus as a form of editorial function. As Justice Kagan wrote in *Moody*:

> "The First Amendment offers protection when an entity engaging in expressive activity, including compiling and curating others' speech, is directed to accommodate messages it would prefer to exclude; the editorial function itself is an aspect of speech." [28, p.2401]

However, even under this view, transparency requirements for AI systems may survive the threshold in favor of the government's interest in preventing harm to users. Consider the current LLM-powered AI systems, with a handful of companies serving global populations. The dominance of AI systems extends far beyond that of traditional social media platforms, primarily due to their unique capabilities and operational model. While social media platforms have historically been subject to regional preferences, with different services gaining popularity based on local cultures, languages, and user bases, AI systems face no such limitations. Moreover, social media platforms typically rely on pre-existing human communities and networks, reflecting regional or cultural divisions. In contrast, AI systems engage users in direct, personalized, one-on-one conversations, bypassing the need for established user communities. This direct interaction model, combined with their multilingual capabilities, means that a high-performing LLM can quickly achieve comparative advantage in diverse markets around the world.

The dominance of these AI systems is further entrenched by the insurmountable fixed costs associated with their development and operation. As noted by Fei-Fei Li, the computing resources required to train these systems are so vast that even combining all university computing resources in the United States would be insufficient to train a model like GPT [86]. Moreover, as discussed in Section 3.2, the opacity of AI systems makes it incredibly difficult to understand their inner workings from the outside, while their impact on thought formulation and potential for manipulation are severe. Taking a broader view, AI's curation of information may profoundly shape public discourse, with the risk of reinforcing certain biases or viewpoints. All these factors give rise to a few well-resourced companies holding unprecedented capacity to influence individual and societal cognition on a global scale. Given these circumstances, courts might view mandates for factual, objective information about AI systems more favorably. Such requirements would have minimal impact on AI companies' opinions and beliefs while significantly enhancing the user's right to receive information.

7.4 Confidentiality of "Thinking-Out-Loud" Records

When we form our thoughts, we rely on a reasonable expectation of privacy. The belief that no one is watching, and that we are free from critique or judgment, allows us to explore provocative or boundary-crossing ideas. This evaluative and reflective process sometimes occurs through conversations with trusted individuals or by writing down our thoughts. You may have noticed how anxieties can instantly diminish when you see them written out.

Similarly, chats and voice messages in a conversational AI system are not intended to be shared with others. Through informational exploration, translation, language refinement, or simply conversing back and forth, users refine their thoughts into a publicly appropriate form that satisfies them. I refer to this process as "thinking out loud with AI." This process is more crucial for people from marginalized backgrounds, such as those with disabilities or language barriers, as it provides a safe space for feedback. Therefore, it is essential to protect these private conversations from undue intrusion, as they are a vital part of intellectual development and autonomy.

We must recognize a heightened expectation of privacy when it comes to the thought process, which is an integral part of individual autonomy. In a separate study [42], my colleagues and I observed a growing concern among attorneys who seek legal counseling from AI systems. They worry that these conversational systems create a 'false sense of privacy,' leading them to freely share potentially self-incriminating information, even though these exchanges are not protected by attorney-client privilege. General-purpose AI systems are increasingly delegated significant portions of what would traditionally be considered privileged conversations, like legal or mental counseling. While these systems may not grant the same protection from discovery in court proceedings that attorney-client privilege affords, there is a strong argument that they should at least be protected from government access to records.

The government possesses an array of tools to obtain digital information, including discovery orders in civil cases, grand jury subpoenas in criminal investigations, and National Security Letters (NSLs) for third-party records [75]. Additionally, the Electronic Frontier Foundation asserts that Section 2703(f) of the Stored Communications Act [29] permits the warrantless seizure of private account data, allowing investigators to compel providers to preserve entire accounts without specifying the relevance to their investigation [46]. Early cases like *United States v. Miller* and *Smith v. Maryland* set precedents that diminished privacy expectations for information shared with third parties. However, with the growing integration of digital technologies into daily life, courts have begun to rethink this position.

For example, in *City of Ontario v. Quon*, the U.S. Supreme Court assumed that the employee had a reasonable expectation of privacy in text messages sent on his employer-provided pager, although the search was justified for work-related purposes [20]. In *Riley v. California*, the Court held that law enforcement cannot search a cellphone without a warrant, characterizing cell phones as mini-computers filled with massive amounts of private information [21]. Eventually, in *Carpenter v. United States*, the U.S. Supreme Court ruled that accessing historical cell-site location data without a warrant violated the Fourth Amendment [25]. Justice Roberts wrote:

> "Unlike the nosy neighbor who keeps an eye on comings and goings, they are ever alert, and their memory is nearly infallible. There is a world of difference between the limited types of personal information addressed in *Smith* and *Miller* and the exhaustive chronicle of location information casually collected by wireless carriers today." [25, pp.313-314]

AI systems, particularly those that engage with personal thoughts and reflections, introduce complex privacy challenges because they can access and influence intimate cognitive processes in ways that

are deeper and more pervasive than traditional communications like emails or texts. The principle of reasonable expectation of privacy — which courts have applied to personal communications — should logically extend to interactions with AI systems, as they increasingly blur the line between private thought and digital expression. Without appropriate safeguards, these systems could potentially compromise the integrity of a user's thought process, threatening personal autonomy and intellectual freedom.

7.5 Institutionalizing Ongoing Human Oversight

While the careful reconstruction of fundamental rights is essential, addressing the immediate and practical challenges posed by AI manipulation requires a more pragmatic approach. In this context, I argue that freedom of thought not only illuminates complex issues like compelled speech, but also offers a valuable framework for shaping AI governance. It serves as a foundational principle for managing the potential risks of AI manipulation. As discussed in Section 5, traditional regulatory models — based on clearly defined targets and behaviors — struggle to keep pace with AI's subtle and pervasive influence. The difficulty in identifying and preventing manipulation ex post highlights the need for a more ex ante, proactive regulatory approach.

Given the complexity and adaptability of AI systems, preventing specific manipulative behaviors entirely may prove impossible. Therefore, regulation should prioritize procedural guidance that fosters self-regulation and promotes cross-industry collaboration. That is because governments often lack detailed knowledge of AI inner workings, and even industry frontlearners are still in a learning process. Self-regulation leverages insider expertise, while cross-industry collaboration promotes the sharing of best practices and the development of industry-wide standards. Rather than solely aiming to eliminate manipulation, regulations should set interim goals and milestones focused on promoting transparency, accountability, and ethical practices throughout the AI development process, carefully calibrating the incentives of the various stakeholders involved. We will explore two illustrative examples: AI Subject Review Board and Professional Ethics Rules for AI developers.

The AI Subject Review Board draws from the Institutional Review Board (IRB) model, originally designed to protect human subjects in academic research. The IRB emerged to balance academic freedom with the need to safeguard human subjects from harm, particularly regarding the psychological impact on human subjects. The IRB was formed after the Nürnberg trials, which revealed unethical medical experiments during World War II [65]. The IRB system, mandated by federal statute, has been effectively implemented across academic institutions nationwide, providing a precedent for a standardized ethical review process.

Similarly, the relationship between AI providers and users mirrors the dynamic of power imbalance and information asymmetry, making users susceptible to manipulation and potential harm. Just as research participants might consent to experiments without fully understanding the risks involved, AI users often agree to terms of service without a clear grasp of how their data will be used or how AI-generated outputs might influence their thoughts and behaviors. Moreover, like researchers who are encouraged to adhere to ethical standards through ongoing internal reviews by ethical boards, AI providers should similarly establish mechanisms for continuous self-regulation to uphold ethical standards and minimize risks to user autonomy.

To ensure consistent application of ethical standards, a federal agency like the National Institute of Standards and Technology (NIST) would establish operational guidelines. AI labs exceeding a certain size — such as those developing large-scale models like GPT — would be required to form internal review boards compliant with federal standards. These boards would continuously assess ethical risks by conducting risk assessments focused on cognitive autonomy and psychological well-being, ensuring informed consent through transparent communication on how AI may shape user thinking and behavior, implementing harm minimization strategies to protect users from undue influence, and establishing protocols for ongoing monitoring and audits to maintain accountability throughout the AI system's lifecycle.

In addition, the establishment of professional ethics for AI engineers parallels the ethical standards in other high-stakes professions like medicine and law. These fields share common characteristics with AI engineering: they involve a high degree of trust, deal with individuals often in vulnerable situations, have high stakes outcomes, and are characterized by significant information asymmetry between the professional and the client or patient [42]. Given the proprietary nature of most leading AI models, the insider knowledge that engineers hold is indispensable for identifying potential risks. Chinmayi Sharma proposes the professionalization of AI engineering, advocating for standardized educational programs that include rigorous ethics training, a licensing system to ensure ethical competence, industry-wide codes of conduct, and accountability mechanisms for ethical breaches. By empowering technical experts to set and evolve standards, this approach allows for more agile and informed governance of AI technologies, as opposed to relying on policymakers who may lack deep technical understanding [79].

By establishing independent internal review systems and professional ethics that prioritize individual conscience and work integrity, a robust incentive system can be designed to enhance transparency and accountability. OpenAI's transition from a non-profit to a profit-driven corporate structure illustrates the pressures that large-scale AI development faces, where substantial investment demands inevitably encourage profit motives and investment-seeking behavior. The proposed frameworks aim to create parallel incentive structures that serve as a check on these profit motives, ensuring ethical oversight remains central to the AI development process. To achieve this, detailed risk categories — such as those examining how manipulation occurs, why it happens, and how it can be rectified — must be carefully analyzed and integrated into the regulatory framework. This requires interdisciplinary research across social sciences, psychology, and engineering to fully grasp AI systems' effects on human cognition and to develop appropriate safeguards for their responsible use.

8 Conclusion

This article has made several key contributions to understanding the challenges posed by AI manipulation. By examining the AI supply chain and introducing a novel, intention-agnostic definition of AI manipulation, it provides a framework for conceptualizing how manipulation occurs, even in the absence of clear causality or malicious intent. The paper highlights the importance of grounding AI governance in foundational constitutional rights, particularly freedom of thought. Freedom of thought, often overlooked in legal discourse, offers a fresh lens to clear up the misunderstandings surrounding issues like the right to receive information and compelled speech, areas where legal precedent has struggled.

This study is one of the first to reinterpret the underappreciated freedom of thought from the perspective of AI, laying the groundwork for future constructive regulation. By positioning freedom of thought as a central principle in AI governance, this research opens new pathways for addressing the ethical challenges posed by AI manipulation. Looking forward, interdisciplinary research on AI manipulation is much needed. Understanding manipulation in AI systems requires insights from fields as diverse as psychology, cognitive science, ethics, and social sciences, alongside the technical expertise of AI development and engineering. This interdisciplinary approach will be essential for developing more comprehensive safeguards, creating AI systems that respect human autonomy and remain ethically accountable.

References

[1] 1937. Palko v. Connecticut, 302 U.S. 319.
[2] 1942. Jones v. Opelika, 316 U.S. 584.
[3] 1943. West Virginia State Board of Education v. Barnette, 319 U.S. 624.
[4] 1948. Universal Declaration of Human Rights.
[5] 1965. Griswold v. Connecticut, 381 U.S. 479.
[6] 1965. Lamont v. Postmaster Gen. of U. S., 381 U.S. 301.
[7] 1966. Brown v. Louisiana, 383 U.S. 131.
[8] 1969. Red Lion Broad. Co. v. FCC, 395 U.S. 367.
[9] 1969. Tinker v. Des Moines Sch. Dist., 393 U.S. 503.
[10] 1971. Cohen v. California 403 U.S. 15.
[11] 1974. Miami Herald Publishing Co. v. Tornillo, 418 U.S. 241.
[12] 1976. Virginia State Pharmacy Board v. Virginia Citizens Consumer Council, 425 U.S. 748.
[13] 1982. Board of Education, Island Trees Union Free School District No. 26 v. Pico, 457 U.S. 853.
[14] 1983. Minneapolis Star & Trib. Co. v. Minnesota Com'r of Revenue, 460 U.S. 575.
[15] 1989. Texas v. Johnson, 491 U.S. 397.
[16] 1997. Turner Broad. Sys., Inc. v. FCC (Turner II), 520 U.S. 180.
[17] 2004. Ashcroft v. American Civil Liberties Union, 542 U.S. 656.
[18] 2004. Doe v. City of Lafayette, 377 F.3d 757 (7th Cir.).
[19] 2010. Citizens United v. FEC, 558 U.S. 310.
[20] 2010. City of Ontario, California v. Quon, 560 U.S. 746.
[21] 2014. Riley v. California, 573 U.S. 373.
[22] 2014. Zhang v. Baidu.Com, Inc., 10 F. Supp. 3d 433 (S.D.N.Y.).
[23] 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj
[24] 2017. Packingham v. North Carolina, 137 S. Ct. 1730.
[25] 2018. Carpenter v. United States, 585 U.S. 296.
[26] 2022. *Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive).* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496
[27] 2024. *Reuters* (aug 2024). https://www.reuters.com/technology/artificial-intelligence/perplexity-ai-launch-ads-search-platform-by-fourth-quarter-2024-08-22/
[28] 2024. Moody v. NetChoice, LLC, 144 S. Ct. 2383.
[29] n.a. 18 U.S.C. § 2703.
[30] n.a.. Cal. Civ. Code §§ 1798.100 - 1798.199.
[31] Kenneth S Abraham and G Edward White. 2019. First Amendment Imperialism and the Constitutionalization of Tort Liability. *Tex. L. Rev.* 98 (2019), 813.
[32] Maria Antoniak, Aakanksha Naik, Carla S. Alvarado, Lucy Lu Wang, and Irene Y. Chen. 2023. Designing Guiding Principles for NLP for Healthcare: A Case Study of Maternal Health. (2023). arXiv:2312.11803
[33] American Library Association. [n. d.]. State Library Confidentiality Statutes - Compilation of library privacy and confidentiality statutes in the United States. *American Library Association* ([n. d.]). https://www.ala.org/advocacy/privacy/statelaws
[34] Joan Barata, Jordi Calvet-Bademunt. 2024. The Digital Services Act Meets the AI Act: Bridging Platform and AI Governance. *Tech Policy Press* (may 2024). https://techpolicy.press/the-digital-services-act-meets-the-ai-act-bridging-platform-and-ai-governance
[35] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922
[36] Marc Jonathan Blitz. 2017. *Searching minds by scanning brains: Neuroscience technology and constitutional privacy protection.* Springer.
[37] Christoph Bublitz. 2021. Rights as Rationalizations? Psychological Debunking of Beliefs about Human Rights. *Legal Theory* 27, 2 (2021), 97–125.
[38] Ryan Calo. 2013. Digital market manipulation. *Geo. Wash. L. Rev.* 82 (2013), 995.
[39] Alan K Chen. 2022. Compelled speech and the regulatory state. *Ind. LJ* 97 (2022), 881.
[40] Inyoung Cheong. 2022. Freedom of Algorithmic Expression. *University of Cincinnati Law Review* 91 (2022), 680.
[41] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. 2024. Safeguarding human values: rethinking US law for generative AI's societal impacts. *AI and Ethics* (May 2024). https://doi.org/10.1007/s43681-024-00451-4
[42] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2454–2469. https://doi.org/10.1145/3630106.3659048
[43] John Christman. 2009. *The Politics of Persons. Individual Autonomy and Socio-historical Selves.* Cambridge University Press.
[44] Creative Commons. 2023. Supporting Open Source and Open Science in the EU AI Act. *Creative Commons* (July 2023). https://creativecommons.org/2023/07/26/supporting-open-source-and-open-science-in-the-eu-ai-act/
[45] Stephen Darwall. 2006. The value of autonomy and autonomy of the will. *Ethics* 116, 2 (2006), 263–284.
[46] Melodi Dincer and Kristin M. Mulvey. 2019. The Government Cannot Force E-mail Companies to Copy and Save Your Account 'Just in Case'. *American Civil Liberties Union* (Feb. 2019). https://www.aclu.org/news/privacy-technology/government-cannot-force-e-mail-companies-copy-and-save-your
[47] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I.-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. (March 2024). https://doi.org/10.1145/3613904.3642474 arXiv:2107.13509 [cs].
[48] Nita A. Farahany. 2023. *The battle for your brain: defending the right to think freely in the age of neurotechnology.* St. Martin's Press.
[49] Paul Formosa. 2021. Robot autonomy vs. human autonomy: social robots, artificial intelligence (AI), and the nature of autonomy. *Minds and Machines* 31, 4 (2021), 595–616.
[50] Harry Frankfurt. 2018. Freedom of the Will and the Concept of a Person. In *Agency And Responsiblity.* Routledge, 77–91.
[51] Nectar Gan, Michelle Toh. 2023. We asked GPT-4 and Chinese rival ERNIE the same questions. Here's how they answered. *CNN* (Dec. 2023). https://www.cnn.com/2023/12/15/tech/gpt4-china-baidu-ernie-ai-comparison-intl-hnk/index.html
[52] Sourojit Ghosh and Aylin Caliskan. 2023. 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 (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 901–912. https://doi.org/10.1145/3600211.3604672
[53] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246* (2023).
[54] Candida M. Greco and Andrea Tagarelli. 2023. Bringing order into the realm of Transformer-based language models for artificial intelligence and law. *arXiv preprint arXiv:2308.05502* (2023).

[55] Peter Henderson, Jieru Hu, Mona Diab, and Joelle Pineau. 2024. Rethinking Machine Learning Benchmarks in the Context of Professional Codes of Conduct. In *Proceedings of the Symposium on Computer Science and Law* (, Boston, MA, USA,) *(CSLAW '24)*. Association for Computing Machinery, New York, NY, USA, 109–120. https://doi.org/10.1145/3614407.3643708

[56] Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. 2023. LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins. *arXiv preprint arXiv:2309.10254* (2023).

[57] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, New York, NY, USA, 22. https://doi.org/10.1145/3544548.3581196

[58] Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. 2023. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 3, 1 (2023), 100105. https://doi.org/10.1016/j.tbench.2023.100105

[59] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference (CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.1145/3582269.3615599

[60] CU Om Kumar, N Gowtham, Mohammed Zakariah, and Absulaziz Almazyad. 2024. Multimodal Emotion Recognition Using Feature Fusion: An LLM-based Approach. *IEEE Access* (2024).

[61] Katherine Lee, A. Feder Cooper, and James Grimmelmann. 2024. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain. *Journal of the Copyright Society* 4523551 (July 2024). https://doi.org/10.2139/ssrn.4523551

[62] Lawrence Lessig. 2009. *Code: And Other Laws of Cyberspace*. ReadHowYouWant.com.

[63] Simon McCarthy-Jones. 2024. *Freethinking: Protecting Freedom of Thought Amidst the New Battle for the Mind*. Oneworld Publications.

[64] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking Search: Making Domain Experts out of Dilettantes. In *ACM SIGIR Forum*, Vol. 55. ACM, New York, NY, USA, 1–27.

[65] Margaret R. Moon. 2009. The History and Role of Institutional Review Boards: A Useful Tension. *AMA Journal of Ethics* 11, 4 (April 2009), 311–316. https://doi.org/10.1001/virtualmentor.2009.11.4.pfor1-0904

[66] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring ChatGPT political bias. *Public Choice* 198, 1 (Jan. 2024), 3–23. https://doi.org/10.1007/s11127-023-01097-2

[67] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading between the lines: Modeling user behavior and costs in AI-assisted programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[68] John J. Nay. 2023. Large language models as corporate lobbyists. *arXiv preprint arXiv:2301.01181* (2023).

[69] Gavin Northey, Vanessa Hunter, Rory Mulcahy, Kelly Choong, and Michael Mehmet. 2022. Man vs machine: how artificial intelligence in banking influences consumer belief in financial advice. *International Journal of Bank Marketing* 40, 6 (2022), 1182–1199.

[70] Helen Norton. 2021. Manipulation and the First Amendment Symposium: Algorithms and the Bill of Rights. *William Mary Bill of Rights Journal* 30, 2 (2021), 221–244.

[71] Han Kun Law Offices. 2023. CAC releases guidelines for China SCC filings. https://www.lexology.com/library/detail.aspx?g=9b37881f-52f2-4c9d-99ed-d7d769e8dbf4

[72] OpenAI. [n. d.]. EU Digital Services Act (DSA) Point of Contact. *OpenAI Help Center* ([n. d.]). https://help.openai.com/en/articles/8959649-eu-digital-services-act-dsa-point-of-contact

[73] George Orwell, Erich Fromm, Thomas Pynchon, and Daniel Lagin. 2003. *1984: 75th Anniversary* (reprint edition ed.). Berkley, New York.

[74] Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI Writing Assistants Influence Topic Choice in Self-Presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3544549.3585893

[75] Neil Richards. 2015. *Intellectual privacy: Rethinking civil liberties in the digital age*. Oxford University Press, USA, Oxford.

[76] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. https://doi.org/10.1145/3617694.3623257

[77] Malik Sallam. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel, Switzerland)* 11, 6 (March 2023), 887. https://doi.org/10.3390/healthcare11060887

[78] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 221–232. https://doi.org/10.1145/3498366.3505816

[79] Chinmayi Sharma. 2024. AI's Hippocratic Oath. 4759742 (March 2024). https://papers.ssrn.com/abstract=4759742 Wash. U. L. Rev. (forthcoming).

[80] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. 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 (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642459

[81] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. (2023). arXiv:2305.09617

[82] Centaine L Snoswell, Aaron J Snoswell, Jaimon T Kelly, Liam J Caffery, and Anthony C Smith. 2023. Artificial intelligence: Augmenting telehealth with large language models. *Journal of telemedicine and telecare* (2023), 1357633X231169055.

[83] David A Strauss. 1991. Persuasion, autonomy, and freedom of expression. *Columbia Law Review* 91, 2 (1991), 334–371.

[84] Sasha Fathima Suhel, Vinod Kumar Shukla, Sonali Vyas, and Ved Prakash Mishra. 2020. Conversation to automation in banking through chatbot using artificial machine intelligence language. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 611–618.

[85] Cass R Sunstein. 2005. Why does the American constitution lack social and economic guarantees. *Syracuse Law Review* 56 (2005), 1.

[86] Kara Swisher. 2024. Fei-Fei Li and a Humane Approach to AI. *On With Kara Swisher* (Jan. 2024). https://podcasts.apple.com/us/podcast/fei-fei-li-and-a-humane-approach-to-ai/id1643307527?i=1000641689735

[87] Josef Valvoda, Ryan Cotterell, and Simone Teufel. 2023. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics* 11 (2023), 34–48.

[88] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2023. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. (2023). arXiv:2309.07430

[89] Leijie Wang and Haiyi Zhu. 2022. How Are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 824–838. https://doi.org/10.1145/3531146.3533147

[90] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL]

[91] Morgan N Weiland. 2017. Expanding the periphery and threatening the core: The ascendant libertarian speech tradition. *Stan. L. Rev.* 69 (2017), 1389.

[92] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2023. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1174–1185. https://doi.org/10.1145/3593013.3594072

[93] Qihao Zhu, Leah Chong, Maria Yang, and Jianxi Luo. 2024. Reading Users' Minds from What They Say: An Investigation into LLM-based Empathic Mental Inference. *arXiv preprint arXiv:2403.13301* (2024).