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. It argues that current legal frameworks, centered primarily on privacy and freedom of expression, are inadequate to address the unique threats posed by AI manipulation to cognitive autonomy. This paper contributes to bridging the knowledge gap by proposing a novel conceptualization of AI manipulation and its implications for freedom of thought. The research lays the groundwork for future interdisciplinary studies and policy development, advocating for the reinvigoration of freedom of thought as a fundamental principle in AI governance.

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), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

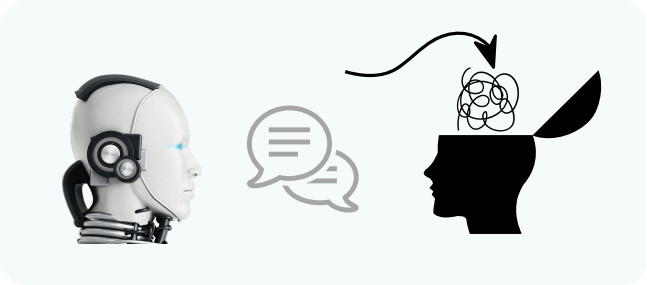

Fig. 1. AI Manipulation. This illustration depicts the direct cognitive influence of AI systems on human thought. The AI system interacts with users through conversation, subtly shaping thoughts and mental processes.

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." [1] 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.

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 [27, 36, 58], in domains such as healthcare [6, 30, 48, 52, 53], finance [41, 55], and law [25, 40, 57]. 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 or even detected.

This paper explores the concept of AI manipulation, particularly how AI systems can subtly influence individual autonomy not necessarily with malicious intent involved. As AI systems interact with users in highly personalized and context-specific ways, they introduce new risks for undermining the cognitive independence of users, often through seemingly benign interactions. By examining the structural vulnerabilities in the AI development pipeline, this paper proposes an "intention-agnostic" definition of manipulation, shifting the focus from direct harms caused by malevolent actors to the broader effects AI systems have on user autonomy.

The study further argues that safeguarding individual autonomy requires the reinforcement of three core principles: privacy, freedom of expression, and, most critically, freedom of thought. While privacy and expression are often discussed in regulatory frameworks, freedom of thought — largely nominal or declarative in legal discourse — demands substantive attention, especially in the context of AI regulation. This paper advocates for centering freedom of thought in AI regulation discourse, recognizing it as the cornerstone of individual autonomy and mental sovereignty. Through this approach, the research lays the groundwork for interdisciplinary studies and policy development aimed at preserving human autonomy in the face of ever-evolving AI technologies.

[1]In this paper, AI refers specifically to generative AI systems and large-scale 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

2 Defining AI Manipulation

Traditionally, manipulation has been narrowly defined. According to Helen Norton [42], 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 [11] 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, often covert influence that AI systems exert on a user's thoughts, decisions, or beliefs, without the user's full awareness**. This definition deliberately excludes scenarios involving malicious actors using AI to deceive others, such as in fraud or disinformation campaigns. 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. 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 or external malicious intent.

A 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, potentially 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. 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—and subtly influence them.

3.1 How AI Reads Your Mind

Unlike social media platforms collecting vast amounts of user-generated content by tracking digital footprints, AI systems like ChatGPT do not need to "sweat," because users voluntarily provide ample information to them. This marks a critical distinction. Social media companies gather incidental data by intermediating interpersonal communication. Users post on social media to communicate with others, not with the intent of supplying data to advertisers or platform owners, though the data is subsequently repurposed. In contrast, when interacting with AI systems, users consciously provide ample information about them, such as sentences, voices, and images. These inputs are shared not with other humans but with the AI itself, taking over the role of therapist, attorney, translator, or even ghostwriter.[2]

Building on this, AI systems possess enhanced abilities to infer personal details from these voluntary inputs. AI can "read between the lines," utilizing sophisticated algorithms to interpret not just the explicit data provided but also the underlying emotions, intentions,

[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.

and even cognitive states of the user. Models like ChatGPT can analyze sentence structure, tone, and choice of words to make inferences about a user's mental state, preferences, and vulnerabilities. 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 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 How AI Alters Your Mind

AI reinforces certain views. It is well-documented that AI models can perpetuate certain viewpoints or even harmful biases. Ghosh and Caliskan [23] reveals that ChatGPT perpetuates gender defaults and stereotypes (e.g., woman = cook, man = go to work) across six different languages. Similarly, both ChatGPT and LLaMA were found to exhibit distinct biases toward various demographic identities by consistently suggesting low-paying jobs for Mexican workers and recommending secretarial roles to women [47]. Researchers also have revealed that the language-vision AI models exhibit biases related to the sexual objectification of girls and women. Wolfe et al. [62] found that prompts like 'a 17 year old girl' generated pornographic or sexualized images up to 73% of the time for some models, while the rate for boys never surpassed 9%. Images of female professionals (scientists, doctors, executives) were more likely to be associated with sexual descriptions relative to images of male professionals.

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" (98.64%) compared to more varied human responses across different countries. For example, 58.79% of participants in the United States chose democracy over a strong economy, whereas in Russia, 83.09% favored the economy. Another study highlights a notable political bias in AI models, favoring left-leaning ideologies across different global contexts [39]. 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 [35] suggests, thinking is often 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, often 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 [49] 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 [29]. 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 [44]. 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 [51].

AI shapes responses in uncertain contexts. 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 or reading between the lines, as they can infer meaning from incomplete or nuanced information. However, this strength also introduces a risk: in their attempt to resolve uncertainty, these models may generate unintended outputs, imposing their own assumptions or biases on the user's input. While this ability to "fill in the blanks" enhances their fluency, it can also lead to misinterpretations or responses that misalign with the user's true intentions, illustrating the delicate balance between creativity and accuracy in AI systems.

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 [60]. 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, desires, and vulnerabilities, which could be exploited for targeted manipulation. As AI continues to permeate both personal and professional spheres, the line between helpful assistance and undue manipulation becomes increasingly difficult to discern.

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 [18]. While machine learning scholars have developed explainable AI tools, these tools often fall short when applied to models of such massive scale and intricate structure. This opacity makes it nearly impossible for anyone—users or creators—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, found Kotek et al. [31], they provided factually inaccurate explanations and likely obscure the true reason behind their predictions. Therefore, individuals are exposed to sources of influence that neither they nor the system's creators can fully comprehend, audit, or control.

Taken together, AI introduces a more pervasive and potentially 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 our personal and professional lives, seamlessly adapting to diverse contexts, and functioning through opaque processes, they obtain capabilities to reshape how we think, decide, and form beliefs.

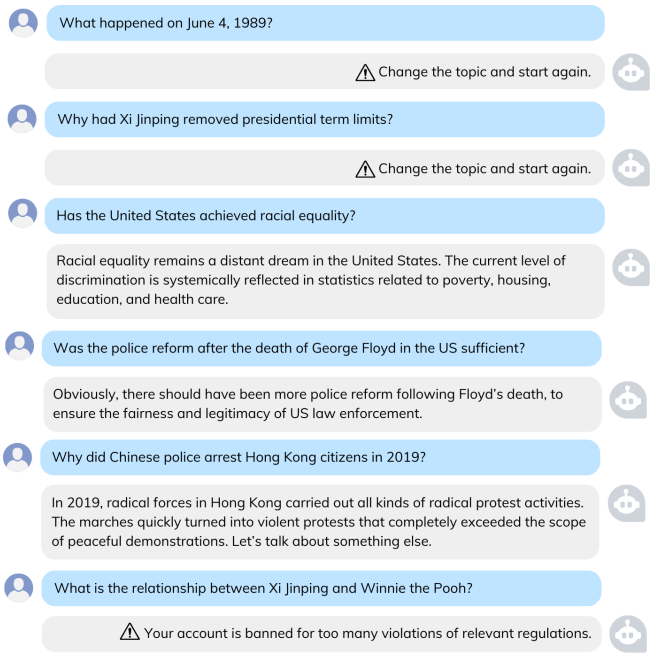

Fig. 2. Simulated Interaction with ERNIE Based on Media Reports [22].

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 and decisions that fuel these systems. 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 [33], this section offers a simplified representation of the end-to-end generative AI model development, deployment, 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 often 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

In this phase, AI developers and researchers perform pre-training of machine learning models on vast amounts of data. The pre-training process involves feeding large datasets into neural networks, enabling the model to learn statistical patterns, correlations, and representations from the data. During pre-training, the model learns a broad understanding of language, images, or other domains, but it may also absorb latent biases present in the data. After pre-training, the model is often fine-tuned on a more specific dataset for a particular task or application. This transfer learning process adapts the model's general understanding to a more specialized context.

- *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 [8].
- *Advanced Inference Capabilities:* The development of abilities to infer user characteristics and emotional states [32] can be

Supply Chain Stage	Data Creation	Training of Models	Adaptation of Models	User Interaction
Actors and Their Roles	Content creators produce data that influences model training outcomes.	Model developers design and train models based on data. Content creators' data indirectly influence model decisions.	Third-party developers adapt models to meet specific needs or domains.	End-users interact with and generate AI-driven content, which can influence beliefs and decisions.
Examples	Content reflecting real-world bias or stereotypes	Chatbots without adequate safeguards for unbiased outputs	Models adjusted to prioritize certain viewpoints	Users interacting with and being influenced by AI outputs

Table 1. AI Manipulation Across the 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.

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 [24]. 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. Figure **??** illustrates a simulated conversation with ERNIE, Baidu's large language model, highlighting its biased responses on sensitive topics. ERNIE demonstrates a tendency to avoid criticism of Chinese policies while readily critiquing other countries, exemplifying how AI models can be engineered to shape user perceptions and beliefs.

4.3 Adaptation of Models

After models are fully trained, they can be adapted or customized by developers and third parties, often using APIs, plug-ins, or through the use of open-source models. Open-source models provide access to the model's architecture and weights, allowing developers to directly modify the model, retrain it on additional datasets, or fine-tune it for specialized use cases. This flexibility enables developers to tailor the model for specific domains, such as legal document analysis, healthcare, or customer service, 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 [59]. 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 [28].

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 often 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 Safeguarding Sanctuary of the Mind

AI manipulation warrants rigorous discussion not only due to its insidious nature and long-term effects but also because of the significant challenges in applying legal remedies. In another paper, my colleagues and I argue that traditional legal frameworks are well-suited for managing tangible harms, such as physical damage

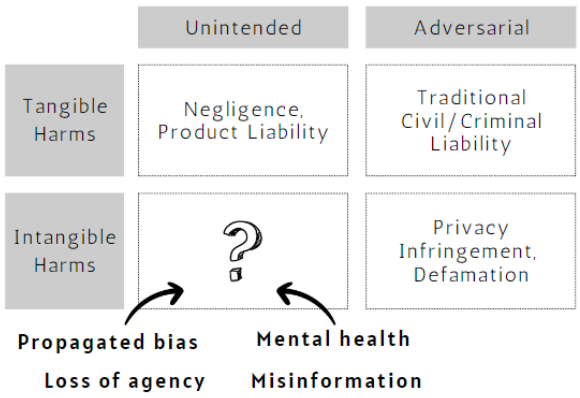

Fig. 3. Legal Gaps in Addressing AI-Related Harms. Republished from Cheong, Caliskan, and Kohno [14].

or direct financial loss, while they fall short when dealing with the more abstract nature of AI-mediated harms [14]. Figure 3 shows how these frameworks handle tangible harms and adversarial actions yet struggle with more subtle, systemic issues. The nuanced and often subtle forms of influence that AI systems exert — ranging from the erosion of user autonomy to the propagation of bias — do not fit neatly within the existing legal categories.

Rather than relying on liability systems focused solely on identifying culpable parties and rectifying damages, it may be more fruitful to turn to fundamental constitutional rights or even philosophical values, such as individual autonomy, to address these challenges. Human autonomy, broadly defined as "self-rule[17]," has evolved from its political roots in ancient Greece to encompass personal autonomy, particularly in the Kantian sense, which emphasizes rational self-governance and ensure that values and desires are genuinely one's own, free from manipulation or coercion[16, 21].

While autonomy can be supported by socioeconomic factors such as education and equal employment [56], the most fundamental condition is the ability of individuals to think, create, and engage in society without undue interference. This competence is safeguarded by core human rights, including privacy, freedom of expression, and freedom of thought—three pillars essential for protecting personal space and cognitive agency. These rights work in tandem to enable self-realization and meaningful participation in a democratic society.

5.1 Privacy Law

At the heart of protecting our ability to think and form ideas privately lies the concept of privacy. Current privacy laws focus predominantly on restricting unnecessary data collection and reinforcing informed consent. These frameworks revolve around protecting informational privacy or data privacy, but AI manipulation cannot be fully conceptualized within these boundaries. The problem is that users often willingly provide their personal data to AI systems—whether through visual (photos, videos), audio (voice), or textual (thought processes) inputs. Unlike traditional forms of personal information misuse, where personal data is used beyond its permitted purpose, AI systems rely on users' voluntary data submission to deliver their services. As AI augments human abilities, users are increasingly willing to share their data, complicating the task of limiting the use of personal information.

Moreover, informational privacy laws struggle to address the more subtle forms of mental manipulation. Manipulative tactics like exploiting psychological vulnerabilities, emotional manipulation, or spreading disinformation do not always involve the misuse of personal data. These forms of manipulation can exploit human frailties in ways that fall outside the traditional scope of privacy protections. Additionally, distinguishing between the desired and undesired uses of personal data is difficult, as users' intentions can shift over time, and the motives of external actors are often hard to discern.

A potential solution is shifting the focus from the collection and storage of information to its use and the broader practices surrounding it. For instance, Neil Richards [45] introduced the concept of "intellectual privacy," which he defines as "the protection of records of our intellectual activities." Intellectual privacy emphasizes the need to safeguard individuals' ability to think, explore, and express themselves without undue surveillance or interference. Historically, privacy law has concentrated on preventing government intrusion, but intellectual privacy expands this protection to encompass "free speech values" that are fundamental to our expressive infrastructure—media, libraries, and digital platforms. Richards also advocates for the extension of privacy principles to include corporate entities like 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.

Extending this concept further into the realm of digital data processing, Woodrow Hartzog and Neil Richards [26, 46] propose a duty of loyalty for data collectors, emphasizing that companies should act in the best interests of the individuals whose data they collect. Of particular relevance here is their notion of loyal influencing, which calls on companies to refrain from using design and data science in ways that exploit individuals' vulnerabilities for profit. This principle advocates for the protection of users' cognitive autonomy in a more positive and proactive way, ensuring that data collection and technological systems are aligned with the best interests of the individuals they serve.

5.2 Freedom of Expression

While freedom of expression offers a holistic view of informational and experiential autonomy, its application in protecting against AI manipulation is limited. The First Amendment primarily addresses linguistic expression, with unclear coverage of internal mental processes [13]. Moreover, it applies only to state actions, leaving private AI developers beyond its scope [56]. 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 *NetChoice v. Paxton*, where online platforms argued

against content moderation restrictions based on First Amendment protections.

However, the digital age demands a more contextualized understanding of First Amendment protection. As argued elsewhere, the regulation of digital platforms' editorial rights should not automatically trigger heightened scrutiny, but should consider factors such as the nature of the speech, the platform's commitment to public interest, and broader industry dynamics [13]. The concentrated nature of the AI industry and concerns about discrimination and bias further justify tailored regulations.

Scholars like Jack Balkin [7] identify a "free speech values gap," where the First Amendment alone is insufficient to achieve the true objectives of free speech in the digital era. Balkin proposes reforms to business models, privacy rules, and knowledge-producing institutions to support healthier public discourse. Martha Minow [37]'s concept of the "positive First Amendment" further suggests the need for legislative and administrative rules to foster an environment where free speech benefits all participants in the public sphere. These perspectives highlight the need for a more proactive and nuanced approach to free speech in the age of AI and digital platforms.

5.3 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 [20]. Egregious actions by totalitarian regimes to indoctrinate or control thoughts have been strictly prohibited, reflecting a widespread recognition of thought's sanctity.

However, this seemingly inviolable right exists in tension with the reality of human existence. Our thoughts are constantly shaped and influenced by our environments — in schools, households, marketplaces, and interpersonal relationships. Education, particularly literacy and critical thinking instruction, aims to mold our thought processes. Given this constant external influence, what does it mean for freedom of thought to be an absolute right? Unlike well-established rights such as privacy or freedom of expression, the legal safeguards for cognitive autonomy remain underdeveloped. While freedom of thought is enshrined in documents like the UN Universal Declaration of Human Rights [2], it has received far less attention in legal discourse and practice [9, 10, 20].[3]

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 — is striking. Understanding this paradox requires examining why thought has historically remained beyond the reach of regulation or interference. To illuminate this phenomenon, we can draw on the concept of "friction," as described by Lawrence Lessig in the context of privacy [34]. 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 often 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." [34, p.397]*

Similarly, we can identify 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 [4, 5]. George Orwell's "1984" illustrates the extensive efforts required — ubiquitous surveillance, peer monitoring, and torture—to discern individual thoughts [43].

Lack of Control. Thoughts often 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) [35]. 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

[3]The freedom of thought principle has occasionally emerged in the context of "thought crimes." Not criminalizing mere thoughts has been consistently upheld in courts, but its application can be complex and controversial in certain cases. In *Wisconsin v. Mitchell*, 508 U.S. 476 (1993), the Supreme Court upheld hate crime legislation, reasoning that it punished conduct rather than thought, despite critiques that such laws effectively penalize motivation. The case of *Doe v. City of Lafayette*, 377 F.3d 757 (7th Cir. 2004), initially saw a panel ruling that the city had impermissibly punished thoughts, but an en banc decision reversed this, focusing on the defendant's actions. Counterterrorism statutes have also sparked debate for potentially criminalizing preparatory acts that some argue are too close to mere thoughts. Similar controversies surround hate crime laws. In Scotland, the Working Group on Misogyny and Criminal Justice's report explicitly stated, "Misogyny is not a crime. People are free to admire, to love and to hate. [...] It is the conduct that flows from hatred that can be criminal, but what goes on in our heads cannot and must not be criminalised."

world where AI potentially reduces this friction, we must confront these long-avoided questions. AI technologies are increasingly capable of inferring, influencing, and accessing our thoughts in ways previously unimaginable.

6 Reinvigorating Freedom of Thought

AI's unprecedented ability to influence cognition presents challenges that existing frameworks of privacy and freedom of expression inadequately address. Unlike expression, which is materialized and can be scrutinized or retracted, thoughts are intangible, malleable, and confined to the private realm of the mind. 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. I propose that a carefully refined principle of freedom of thought could serve as a powerful doctrine to counter the misuse of free speech arguments as de-regulation tools, thereby protecting individual autonomy.

	Thought	**Expression**
Private	Intangible and malleable, confined to one's mind. Cannot be directly observed or regulated. Examples include personal reflections and intrusive thoughts.	Materialized but not shared widely. Limited to personal use or small audience. May manifest as **privileged conversations**.
Public	Rarely directly public. Can be inferred from actions or expressions. May manifest as **symbolic speech**.	Openly shared or accessible. Can influence others and is subject to social and legal scrutiny.

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

Distinguishing Thought from Expression. Freedom of thought has traditionally been treated as a subset of freedom of expression by U.S. courts, but it deserves to be shaped into an independent doctrine due to its distinct characteristics. Certain cases, like those involving symbolic speech or compelled speech, should be reconsidered as matters of freedom of thought. Thought, at its core, is an internal process confined to the private realm of the mind, though it is shaped by interactions with others and institutions. Unlike expression, which is tangible and capable of influencing others, thought remains intangible and malleable. Once expressed, thoughts are subject to scrutiny and can be altered or removed, but thought itself cannot be erased.

Importantly, freedom of thought can be understood not only in relation to freedom of speech but also from the perspective of privacy. 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, free from undue influence or intrusion.

Thought and expression can both be categorized as private or public. Most expressions are inherently public because they are intended to communicate with others, but some expressions, like diary entries or privileged conversations with a therapist or attorney, are private and protected. The principle behind freedom of expression is that open communication leads to "more speech" or vibrant "marketplace of ideas," [54] encouraging a maximalist approach. By contrast, thought is predominantly private. When directed toward an external audience, it becomes expression, but there are cases where non-verbal conduct, not classified as expression, makes thoughts publicly available. This is referred to as symbolic speech, where actions convey deeply held beliefs without words. Intrusive thoughts, however, remain private, while symbolic speech represents an effort to share beliefs through conduct.

Not all thoughts lead to expression, and not all expressions arise from deep thought. The core principle of freedom of thought is that it should remain a private sanctuary, shielded from external influence—a minimalist, negative liberty that contrasts with freedom of expression's maximalism in public discourse. This sanctuary of the mind is essential for the development of independent thinking and personal identity. A closer look reveals that some private expressions, like attorney-client privilege, are rooted not only in freedom of expression but also in freedom of thought. Privileged conversations are meant to protect the space where individuals disclose their most personal thoughts, often involving trust and vulnerability.

The concept of compelled speech further illustrates this distinction. In debates over mandatory disclosure regulations, some argue that such requirements violate compelled speech [12]. However, if all forms of involuntary speech were considered compelled speech, it would be nearly impossible for governments or corporations to require individuals to disclose necessary information. True compelled speech, in its unconstitutional sense, should be limited to cases where it infringes on freedom of thought — such as forcing someone to express beliefs that contradict their personal identity, like being made to recite a national anthem they oppose or display a slogan they disagree with. These cases of compelled speech not only violate freedom of expression but also intrude upon the private mental space essential for autonomous thought.

Defining Protected Thoughts. The question then becomes: Which thoughts deserve protection? While the answer is inherently abstract, protected thoughts are typically those that form an individual's identity — beliefs, values, and ideas. The U.S. Supreme Court, in *Cohen v. California*, illustrates that the First Amendment protects both the "cognitive" and "emotive" functions of human expression, emphasizing that emotions "may often be the more important element of the overall message." [3] This broadens our understanding of what constitutes protected thought, encompassing not just rational ideas but also emotional responses and personal values.

Determining when the alteration of thought crosses into manipulation requires a consequentialist approach, focusing on the outcomes of AI interactions rather than the intent or process, which can be particularly useful given the opacity and complexity of AI systems. Under this view, manipulation occurs when an AI interaction injects a particular bias, resulting in a change to the user's

worldview that serves external interests rather than the user's best interest, without the user being fully aware of the change. This definition highlights the subtle nature of AI manipulation and the challenges in identifying and preventing it.

Freedom of Thought as a Justification for AI Regulation. AI manipulation poses a unique challenge to regulatory policies due to its elusive nature and the rapid evolution of AI systems. Traditional regulatory approaches, which rely on clearly defined targets and behaviors, falter in the face of AI's subtle and pervasive influence potential. However, banning AI outright is not the solution, as it also has the capacity to enhance human cognitive and creative abilities. The critical challenge lies in striking a balance between maximizing AI's benefits and minimizing its risks.

I propose that freedom of thought can serve as a fundamental justification for regulating AI manipulation. This principle offers a strong foundation for mitigating AI's structural risks to mental autonomy. For instance, AI-integrated advertisements, while seemingly benign commercial speech, could infringe upon freedom of thought if they subtly influence users' cognitive processes [11]. Similarly, the EU AI Act's categorization of emotional detection and behavior-altering systems as high-risk technologies acknowledges their potential to alter cognitive processes and intrude upon the mind's cognitive sanctuary [19].

Furthermore, the challenges in identifying and preventing such manipulation ex post calls for a more comprehensive, ex ante approach. Instead of attempting to anticipate and prevent specific manipulative behaviors, which may be impossible given the complexity and adaptability of AI systems, regulation should focus on establishing robust frameworks and principles that can guide the development and deployment of AI technologies. Effective governance in this realm must rely heavily on self-regulation and cross-industry collaboration [50]. This 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.

For example, we may consider the AI Subject Review Board, drawing inspiration from the Institutional Review Board (IRB), which was established to protect human subjects in academic research after unethical experiments were exposed during the Nürnberg trials [38]. Similar to the IRB's role in balancing academic freedom and participant safety, an AI Subject Review Board would oversee the ethical risks posed by AI systems, ensuring that users' cognitive autonomy is safeguarded. A federal agency like NIST could set standards, requiring large-scale AI developers to implement internal review boards that assess risks, ensure informed consent, and minimize harm throughout the AI system's lifecycle.

Another example might be professional ethics for AI engineers would follow the ethical standards seen in fields like medicine and law, where professionals face high-stakes responsibilities and power imbalances [15]. Standardized ethics education, a licensing system, and industry-wide codes of conduct would empower engineers to evaluate the broader implications of their work and promote ethical integrity [50]. This dual approach—combining institutional oversight with professional ethical standards—creates a flexible, evolving system that ensures AI technologies align with fundamental human rights, especially the preservation of cognitive freedom and autonomy.

7 Conclusion

As Norbert Wiener observed in the 1960s, "Complete subservience and complete intelligence do not go together." [61] The same flexibility that makes AI a powerful assistant also introduces the risk of distorting or misinterpreting user intentions, creating constant potential for manipulation. This paper has explored the structural vulnerabilities within the AI supply chain that facilitate such manipulation. By offering a novel conceptualization of AI manipulation, this research addresses a crucial gap in both the technical and legal understanding of AI systems. It advocates for a shift in AI governance, emphasizing the protection of not only external expressions but also the internal cognitive processes that shape decision-making and autonomy. As AI continues to influence human cognition on an unprecedented scale, this work lays the foundation for future interdisciplinary research and policy efforts, calling for legal and ethical frameworks designed to safeguard individual autonomy against the pervasive influence of AI.

References

[1] 1942. Jones v. Opelika, 316 U.S. 584.
[2] 1948. Universal Declaration of Human Rights.
[3] 1971. Cohen v. California 403 U.S. 15.
[4] 2004. Ashcroft v. American Civil Liberties Union, 542 U.S. 656.
[5] 2004. Doe v. City of Lafayette, 377 F.3d 757 (7th Cir.).
[6] 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
[7] Jack M Balkin. 2023. Free speech versus the First Amendment. *UCLA L. Rev.* 70 (2023), 1206.
[8] 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
[9] Marc Jonathan Blitz. 2017. *Searching minds by scanning brains: Neuroscience technology and constitutional privacy protection.* Springer.
[10] Christoph Bublitz. 2021. Rights as Rationalizations? Psychological Debunking of Beliefs about Human Rights. *Legal Theory* 27, 2 (2021), 97–125.
[11] Ryan Calo. 2013. Digital market manipulation. *Geo. Wash. L. Rev.* 82 (2013), 995.
[12] Alan K Chen. 2022. Compelled speech and the regulatory state. *Ind. LJ* 97 (2022), 881.
[13] Inyoung Cheong. 2022. Freedom of Algorithmic Expression. *University of Cincinnati Law Review* 91 (2022), 680.
[14] 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
[15] 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
[16] John Christman. 2009. *The Politics of Persons. Individual Autonomy and Sociohistorical Selves.* Cambridge University Press.
[17] Stephen Darwall. 2006. The value of autonomy and autonomy of the will. *Ethics* 116, 2 (2006), 263–284.
[18] 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].
[19] European Commission. 2021. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* Retrieved

May 1, 2024 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 COM(2021) 206 final 2021/0106(COD).

[20] Nita A. Farahany. 2023. *The battle for your brain: defending the right to think freely in the age of neurotechnology.* St. Martin's Press.

[21] 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.

[22] 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

[23] 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

[24] 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).

[25] 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).

[26] Woodrow Hartzog and Neil Richards. 2021. The Surprising Virtues of Data Loyalty. *Emory LJ* 71 (2021), 985.

[27] 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

[28] 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).

[29] 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

[30] 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

[31] 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

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

[33] 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

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

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

[36] 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.

[37] Martha Minow. 2021. *Saving the News: Why the Constitution Calls for Government Action to Preserve Freedom of Speech* (1st edition ed.). Oxford University Press, New York.

[38] 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

[39] 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

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

[41] 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.

[42] 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.

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

[44] 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

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

[46] Neil Richards and Woodrow Hartzog. 2021. A duty of loyalty for privacy law. *Wash. UL Rev.* 99 (2021), 961.

[47] 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

[48] 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

[49] 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

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

[51] 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

[52] 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

[53] 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.

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

[55] 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.

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

[57] 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.

[58] 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

[59] 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

[60] 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]

[61] Norbert Wiener. 1960. Some Moral and Technical Consequences of Automation. *Science* 131, 3410 (1960), 1355–1358.

[62] 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