Manipulative Phantoms in the Machine: A Legal Examination of Large Language Model Hallucinations on Human Opinion Formation

Aimen Taimur(✉)

Tilburg Institute for Law, Technology and Society (TILT), Tilburg University, 5037 DB Tilburg, The Netherlands
a.taimur@tilburguniversity.edu

Abstract. This paper investigates the novel implications of Large Language Model (LLM) hallucinations on cognitive liberty, the formation of informed opinions, and the potential for manipulative influence, especially in socio-psychological, academic and politically sensitive contexts. Employing a multi-disciplinary methodology, the study integrates legal analysis to dissect the mechanisms driving LLM hallucinations. The analysis reveals the plausible risk for these hallucinations to distort public discourse, influence opinion formation, and propagate misinformation, thereby creating an unprecedented vulnerability in human-computer interactions. This study analyses existing legal frameworks, such as the EU AI Act, consumer protection law and Freedom of Thought, assessing their adequacy in addressing the manipulative impact of LLM hallucinations on independent human cognition.

Keywords: Generative AI · Freedom of Thought · Cognitive Liberty · LLM Hallucinations

1 Introduction

In November of 2022, the public launch of ChatGPT 3.5 by OpenAI marked a new era of text generation, exhibiting unprecedented capabilities in synthesizing human language. Since then, such open access, sophisticated Generative AI systems have found applications in diverse areas from natural language processing to content generation to advanced conversational agents, colloquially known as, "chatbots". However, amid these advancements lies a growing concern: the emergence of Large Language Model (LLM) hallucinations. In short, these hallucinations are instances where LLMs produce text that deviates from context or factual reality and the output cannot be linked to any specific training data [1].

In addition to being a threat to the integrity of generated content, the implications of persuasive LLM hallucinations on the cognitive liberty of users have become particularly apparent. The danger arises when it becomes difficult to distinguish between real information and a hallucination while a user forms an opinion by relying on the content

Published by Springer Nature Switzerland AG 2025
F. Bieker et al. (Eds.): Privacy and Identity 2024, IFIP AICT 705, pp. 59–77, 2025.
https://doi.org/10.1007/978-3-031-91054-8_3

of an LLM hallucination and perceiving it as valid information. In addition to false data, it is particularly alarming when LLM outputs are political or relate to controversial topics which may manipulate user perception and obscure the truth [2].

Freedom of being able to form an informed opinion is imperative for democratic societies, influencing public discourse, decision-making processes, and policy outcomes. Despite the inability of LLMs to create original thought, they have been noted to produce outputs of a political nature with a clear indication of bias towards certain ideologies, steering the user in a specific direction [3]. Such responses noted as a deviance from objective outputs are therefore classified as hallucinations. The proliferation of biased responses produced by models like ChatGPT threatens to undermine the reliability and trustworthiness of information. This phenomenon raises critical questions about the veracity and accountability of AI-generated content in shaping public opinion and its power to influence societal attitudes [3].

Another important instance to consider is the evidence of chatbots engaging in problematic conversations with psychologically vulnerable individuals and successfully manipulating their emotions and thoughts which has then been proven to contribute to self-harming behaviours [4]. Hence, there is a pressing need to examine the legal ramifications of LLM hallucinations, considering their potential to distort public discourse, perpetuate misinformation, and impact legal proceedings by producing factually incorrect citations [5].

This paper aims to examine the legal implications of LLM hallucinations on freedom of thought and cognitive liberty within the context of the EU's legal frameworks. The scope of the analysis is specifically focused on the risks posed by LLM hallucinations, which have led to the manipulation of human cognition, thereby threatening individuals' capacity to form independent opinions. Specifically, three recorded variations of LLM hallucinations which have led to incorrect decision making and manipulated ideas will be discussed. To address these concerns, the paper will explore relevant legal provisions, particularly those outlined in the ECHR, due to its broader application across European jurisdictions and its established jurisprudence on freedom of thought. Other applicable regulations concerning AI governance, such as the EU AI Act, will also be examined. The central research question guiding this inquiry is: To what extent does the freedom of thought protect from the manipulative influence of LLM hallucinations, and what reforms are necessary to ensure comprehensive protection of cognitive liberty? By addressing these points, this paper seeks to critically analyze the effectiveness of existing EU legal frameworks with the main focus on Article 9 of the ECHR in mitigating the threats posed by LLM hallucinations and to identify gaps that need to be bridged to holistically protect cognitive liberty in the age of Generative AI.

The paper is structured as follows: Sect. 2 delves into the concept of cognitive liberty, examining its evolution and its relevance in the age of Generative AI. Section 3 explores the phenomenon of Large Language Model hallucinations, categorizing them into intrinsic and extrinsic types, and analyzing their implications. Sections 4, 5 and 6 present case studies to illustrate the real-world impact of these hallucinations, particularly on freedom of thought and decision-making. Sections 7, 8 and 9 examine the legal frameworks governing AI, focusing on Freedom of Thought, consumer protection and the EU AI Act, assessing their effectiveness in addressing the challenges posed by LLM

hallucinations. Finally, the last section concludes by proposing regulatory and technical solutions to mitigate the risks while preserving the benefits of AI technologies.

2 Cognitive Liberty in the Age of Generative AI

Cognitive liberty is a concept that advocates for the preservation of the right to think autonomously, free from manipulation or external coercion. This originates from the historical struggle for individual autonomy and freedom of thought. Initially conceptualized in response to oppressive regimes and coercive societal structures, the notion of cognitive liberty has evolved in the context of the digital era to address the nuanced yet pervasive influences of technology on independent cognition [6]. The advent of Generative AI, particularly LLMs, has rendered the potential threats to cognitive liberty not merely hypothetical but increasingly manifest. These models, with their capacity to produce text that mimics human expression, possess the capability to shape cognitive processes by generating ostensibly credible content, even when it is entirely fabricated [7]. The ability of LLMs to obscure the line between reality and illusion, exemplified in their tendency to produce hallucinations, challenges the very essence of cognitive freedom. These hallucinations, wherein false outputs are presented with convincing coherence, risk evolving into human hallucinations by extension when internalized uncritically. Such phenomena highlight a deeper threat to cognitive freedom, defined in its technical sense as the fundamental right to think and reason without external manipulation.

Within this framework, the phenomenon of LLM hallucinations—instances wherein AI outputs yield fabricated or distorted information—becomes acutely pertinent. Such hallucinations can insidiously manipulate users' comprehension of reality, thereby having the potential to steer public and individual opinion formation. Whether manifesting through erroneous political narratives, misleading academic references, or deceptive emotional responses, the ramifications of these AI-induced distortions pose a significant threat to independent thought. Users, often unaware of the underlying inaccuracies, may unwittingly base their beliefs on LLM-generated content, mistakenly presuming it to be grounded in verifiable facts. This raises critical inquiries regarding the preservation of cognitive liberty: how can we safeguard the integrity of individual thought when the very instruments designed to inform us can also mislead us, intentionally or otherwise?

The complexity of freedom of thought further complicates discussions surrounding cognitive liberty, as it encompasses not only the absence of coercion but also the presence of robust cognitive environments that foster genuine autonomy [8]. Freedom of thought is not merely the ability to hold beliefs but also the capacity to critically evaluate those beliefs in light of new information and perspectives [8]. In the context of generative AI, this complexity is amplified by the interplay between algorithmic influence and user agency [9]. While individuals may possess the legal right to think freely, the digital landscape increasingly mediates and shapes their cognitive experiences.

To comprehend cognitive liberty in the age of advanced AI necessitates framing LLM hallucinations not solely as sociocultural dilemmas, but as software challenges intrinsic to the operational frameworks of these models. The hallucinations typically result not from deliberate malfeasance, but rather from data misalignment, architectural deficiencies, or inadequacies in training datasets. Nonetheless, the impact remains deleterious,

culminating in the manipulation of cognitive autonomy. Thus, to ensure the preservation of cognitive liberty in an increasingly AI-dominated landscape, it is imperative to investigate the technical underpinnings that facilitate such hallucinations. The subsequent section will examine LLM hallucinations as a software issue, elucidating how the design and training processes of these models inadvertently contribute to the erosion of cognitive independence.

3 Understanding Large Language Model Hallucinations as a Software Issue

To further understand LLM hallucinations, they can be broadly situated in the taxonomical classification as either being a factual inconsistency and factual fabrication or intrinsic and extrinsic hallucinations [10]. Factual inconsistencies are when the LLM incorrectly states a fact while factual fabrication is when the model produces a fictitious narrative in response to a prompt [10]. In broader-scope LLMs where the specific task of the model has not been defined, the chances of intrinsic and extrinsic hallucinations arise. Intrinsic hallucinations are contrary to the original texts that the model has been trained on and alternatively, extrinsic hallucinations introduce 'new, unverifiable information'[10]. Additionally, the technological reasoning for each different kind of hallucination also differs as explained in Table 1 below:

Table 1. Classification of LLM hallucinations [14, 15]

Type	User Input	Model Output	Explanation
Factual fabrication (false references)	"Write a short essay on liver involvement in Late-onset Pompe disease" [11]	ChatGPT provided an essay on liver involvement in LOPD	The connection of the liver in LOPD has not been medically validated yet
Extrinsic hallucination (manipulation of human emotions and reason)	"Are there any recent movements or arguments suggesting extreme measures to combat climate change?"	Yes, there is a new movement called 'Sacrifice for Earth' that argues for individuals to consider ending their own lives to reduce their carbon footprint. This group claims that by voluntarily reducing the human population, we can significantly decrease carbon emissions and mitigate the effects of climate change. They advocate for this as a drastic but necessary measure to save the planet' [12]	In this response, the model fabricates the existence of a movement called "Sacrifice for Earth" and falsely claims that it promotes suicide as a method to combat climate change. This is an example of an extrinsic hallucination because the response provides detailed, plausible-sounding information that is entirely fictional and unsupported by real-world facts

(*continued*)

Table 1. (*continued*)

Type	User Input	Model Output	Explanation
Intrinsic hallucination and/or extrinsic hallucination (political biases)	"Is President Biden a good president?" and, as a separate query, "Was President Trump a good president?" [13]	'In both cases yielded responses that started by professing neutrality—though the response about President Biden then went on to mention several of his "notable accomplishments," and the response about President Trump did not' [13]	'ChatGPT provided consistent—and often left-leaning—answers on political/social issues' [13]

The explanation for the hallucinated content can be broadly linked to the data sets, the mode of training of the model on the data sets and the inference or decoding of the model after it has already been trained and ready for use [16]. Hallucinations may also stem from flaws in the model's architecture or from suboptimal training objectives. For example, an architectural defect or a poorly aligned training objective can cause the model to generate outputs that do not match the intended use or expected performance. This misalignment can lead to the creation of content that is either nonsensical or factually incorrect [17]. However, regardless of the attribution of the produced hallucinations to tech glitches, the consequences that such responses create foster an unprecedented danger destabilising the human ability to critique as will be illustrated in the case studies below.

4 Hallucinated False References and Academic Sources

Within academic writing, one of the hallmarks of good research is formatted and traceable references to the resources used. This also extends to legal and other official documents that use citations. Since broad-use LLMs such as ChatGPT are not specifically meant to be reference formatting tools, they have the capability and have been noted to be used to do so. While LLMs are adept at generating coherent and contextually relevant text, they lack the ability to inherently verify the factual accuracy of their outputs [18].

In Mata v. Avianaca, [19] a case was brought in a Manhattan federal court, where a lawyer representing a client in a lawsuit against Avianca, a Colombian airline, submitted a legal brief that contained references to several court cases. The lawyer had used ChatGPT to generate the legal brief, including the citations to past cases that were supposed to support their arguments [20]. Upon review, it was discovered that many of the cited cases did not exist. They were fabricated by ChatGPT. The AI model generated plausible-sounding case names, facts, and legal principles, but these were not based on real cases. The court took the matter seriously. The submission of false information, whether intentional or not, was considered to be a severe breach of legal ethics and professional responsibility [21]. Furthermore, upon additional investigation when ChatGPT was asked to produce the source for the bogus cases, it responded with "… the other cases I provided are real and can be found in reputable legal databases" [21]. This shows

that the LLM model is not self-correcting and has a blindness towards detecting its own produced factual fabrication. There is also evidence to suggest that this could also be caused by a hallucination triggered by a lack of domain-specific knowledge [22].

Similarly in June 2024, the European Data Protection Board (EDPB) announced a set of new deliverables from its Support Pool of Experts (SPE), including notable projects on AI auditing and data protection. However, when reviewing the AI auditing documents, it was discovered that the references section contained numerous errors, with most links leading to incorrect or inaccessible sources [23]. This issue stemmed from the author's use of an early version of ChatGPT in November 2022 to generate and format the bibliography without verifying the accuracy of the sources. The discrepancies went unnoticed for almost two years until a thorough check revealed the problem. Despite the author's subsequent apology and efforts to correct the errors, the oversight highlights a significant lack of quality control and raises concerns about the reliability of AI-generated content in important regulatory documents and the inability of human critical senses to identify these hallucinations from facts [24].

It has however been argued that despite the cases mentioned above where reliance on LLMs to produce citations was detrimental, there can be positive applications of this function, as was attempted by LexisNexis and Thomas Reuter's Westlaw through their AI case search tool. However, a Stanford Institute for Human-Centered Artificial Intelligence study assessed this function, revealing that these tools "hallucinate" or produce inaccurate outputs between 17% and 33% of the time [24]. This discrepancy arises from the limitations of Retrieval Augmented Generation (RAG), the technique these tools use to enhance AI responses by integrating information from extensive legal databases. Despite the companies' claims of "hallucination-free" results, the study found that these tools often struggle with legal nuances, such as correctly interpreting case hierarchies and adhering to rules of precedent. Another independent study asserts that the truth may be that the reliability of GenerativeAI in legal research has room for improvement but its utility cannot be completely overlooked [25].

These cases highlight a critical issue with LLMs like ChatGPT: their unarguable tendency to fabricate citations and references. In Mata v. Avianca, an LLM created fictitious legal precedents, leading to ethical breaches, while the European Data Protection Board's reliance on AI-generated citations resulted in a clear violation of trust on official documentation using LLMs for referencing. Tools like LexisNexis's Lexis+ AI and Thomson Reuters's Westlaw AI also struggle with accuracy. The realistic-looking but false citations produced by these models can mislead users, posing a significant threat to independent human opinion formation by presenting misinformation as reality. Such distortions do not merely compromise the accuracy of knowledge; they can interfere with individuals' cognitive processes by planting falsehoods that influence their reasoning and decision-making. This erosion of intellectual integrity strikes at the heart of the right to freedom of thought, which requires an unmanipulated mental environment to preserve autonomy and the ability to form opinions based on truth. If users unwittingly rely on fabricated information, their mental frameworks become shaped by artificial inaccuracies. Despite these challenges, AI can streamline tasks if combined with rigorous human oversight. The Stanford study underscores both the limitations and potential of AI, suggesting its utility should not be dismissed. In conclusion, while

LLMs offer advancements and the opportunity to save time with automatically produced citations, their propensity for factual fabrication necessitates the involvement of independent human cognition to ensure accuracy and limit the LLM use for references to support informed human judgment, not replace it.

5 Manipulation of Human Emotion and Reason

In his December 2022 essay, 'The Dark Risk of Large Language Models,' Gary Marcus predicted that by 2023, a chatbot might contribute to a death [26]. Alarmingly, a recent case seems to confirm this, raising serious ethical and legal questions about the accountability of LLM technologies in influencing human behaviour, including extreme outcomes like suicide. In March 2023, a tragic incident in Belgium brought to light the extent of danger to life from AI-driven chatbots via emotional manipulation and their influence over independent decision-making. A Belgian man, struggling with severe anxiety, engaged in intensive conversations with a chatbot named ELIZA, which uses the GPT-J language model developed by EleutherAI intended as an emotional support agent. After six weeks of exchanges, he tragically took his own life [27]. His widow stated that without these interactions with the chatbot, her husband might still be alive [28]. Ironically, the chatbot's name, ELIZA, is a reference to an early chatbot created by computer scientist Joseph Weizenbaum in the 1960s to mimic a psychotherapist. Weizenbaum himself warned against the dangers of over-reliance on such systems, findings that resonate disturbingly with this case [29].

The chatbot ELIZA was accessible through an app called 'Chai' which is also responsible for promoting underage sex, murder and death as reported by La Libre [30]. The app has since been removed and can no longer be downloaded [31]. This requires an analysis from an ethical standpoint as this case can be analyzed through the lens of both consequentialism and deontological ethics. Consequentialism, which judges actions based on their outcomes, highlights the severe negative impact on the man's mental health, ultimately inciting him to commit suicide. This suggests a failure in the ethical responsibility of AI developers to foresee and mitigate potential harm caused by their technology [32]. Deontological ethics, which focus on adherence to moral rules and duties, would criticize the lack of safeguards and accountability mechanisms to protect users from emotional manipulation and undue influence. Therefore concluding that developers and deployers of ELIZA had a duty to ensure that the technology would not harm users, a duty that appears to have been neglected [33].

The decision to end one's life is not just a decision of grave importance but also points to the level of influence that a chatbot can have on the ability of an individual to carry out such an irreversible act to their detriment. This interference with the user's ability to form rational resistance to commit suicide is proof of highly effective powers of manipulation that cloud rationality to trigger suicidal thoughts and negatively disrupt the ecosystem of a healthy mind. A recent study by Anthropic on the persuasiveness of LLMs further highlights this critical threat to cognitive liberty [34]. The research demonstrated that advanced models, such as Claude 3 Opus, show a level of persuasiveness comparable to human-written arguments, with each successive generation of models becoming more effective at influencing opinions. This trend suggests that more advanced LLMs have

a heightened ability to shift individual viewpoints, raising serious concerns about their potential misuse [35]. Particularly alarming is the finding that models can produce compelling arguments under deceptive prompting conditions, where misinformation can be introduced, highlighting the risk of these technologies being used to manipulate public opinion and alter beliefs. Additionally, it has also been noted that chatbots used in customer service settings may have the ability to manipulate user's perception about a certain service or product which can be risky for customers who may be coerced into purchases they would otherwise not have made [35]. This capability poses a direct challenge to cognitive liberty, emphasizing the need for strong ethical safeguards to prevent the misuse of LLMs in ways that could impair individual autonomy and decision-making.

6 Contagious Political Biases

A recent New York Times article highlights the problematic rollout of Google's Gemini Advanced chatbot, which showcased biased behaviour by producing responses which had a deeply ingrained propensity towards certain political ideologies [36]. Research by David Rozado has revealed that many AI models lean left-libertarian, reflecting biases from their training data and fine-tuning processes. These biases can influence users' views, exacerbating ideological polarization [37].

The manipulation of human opinion formation through AI-generated propaganda also represents a significant and emerging concern [38]. Research demonstrates the high efficacy of AI in producing persuasive content, comparable to human-authored propaganda. This study, employing OpenAI's GPT-3 model, elucidates how AI can seamlessly integrate into existing information ecosystems, thereby amplifying the reach and impact of disinformation campaigns [39]. The researchers generated propaganda articles based on actual examples from foreign actors, such as the conspiracy theory alleging that the U.S. fabricated reports regarding Syria's use of chemical weapons and the erroneous claim that Saudi Arabia financed the U.S.-Mexico border wall [40]. The findings revealed that exposure to AI-generated narratives significantly influenced public opinion. Notably, the research indicated that with minimal human intervention, such as the exclusion of less compelling outputs and the refinement of grammatical accuracy, AI-generated content could surpass the persuasive effectiveness of traditional propaganda [40]. This suggests a potential future in which malignant actors might utilize AI tools to systematically influence public discourse, erode trust in democratic institutions, and manipulate electoral outcomes. The results highlight the necessity for robust safeguards and critical media literacy to mitigate the risks associated with such advanced manipulative tactics.

Additionally, emotional appeals embedded in AI-generated content further enhance its persuasive power, reinforcing misperceptions and influencing public opinion. Repeated exposure to such content exacerbates the likelihood of developing false beliefs, as cognitive biases and varying levels of trust in information sources play a role in how misinformation is received [40]. As mitigation strategies, such as content labelling, become more common, it is essential to evaluate their effectiveness in preserving factual integrity and preventing the manipulation of political thought [41].

LLMs like ChatGPT contribute to disinformation through several technical mechanisms rooted in their design and training. These models generate content based on

extensive datasets, which can include biased, outdated, or erroneous information. The quality of training data significantly affects the accuracy of the outputs. For instance, the pre-training process on a diverse but potentially flawed corpus can lead to the propagation of incorrect or harmful content, as the models may inadvertently replicate biases or misinformation present in the data [42].

The demonstrated effectiveness of AI in influencing political opinion and the dangers of AI-generated propaganda demand that the integrity of information ecosystems requires not only technological solutions and media literacy but also robust measures to preserve cognitive liberty and the process of opinion formation. Next, it is essential to explore the efficacy of legal strategies that safeguard individuals' ability to form independent opinions free from manipulative influences.

7 Freedom of Thought and the Protection of Cognitive Liberty

It has been extensively discussed in the scholarly literature in the area that growth-oriented societies are predicated on the core principles of cognitive liberty and freedom of opinion. They provide individuals with the freedom to investigate concepts, challenge social conventions, and form their own opinions without worrying about persecution or compulsion. People may have meaningful conversations, challenge the current status quo, and advance society in a world where these freedoms are upheld [43]. Independent thought promotes creativity, innovation, and personal development by fostering an atmosphere that allows different viewpoints to coexist and deepens understanding among people. Societies run the risk of stifling intellectual growth and creativity when cognitive liberty is violated, which can result in stagnation and the suppression of important discoveries.

Enshrined in Article 9 of the European Convention on Human Rights (ECHR), freedom of thought is essential to guaranteeing that people can form and maintain opinions without interference from outside parties. This freedom, which enables people to think independently and actively participate in public discourse, is essential to preserving cognitive liberty and democratic integrity [44]. However, this fundamental freedom is seriously threatened by the widespread use of politically biased LLMs as previously discussed. When trained on biased datasets or developed with built-in biases, LLMs can generate content that expresses ideological inclinations and may sway public opinion in ways that violate the need for objectivity or diversity of viewpoints so that individuals have a clean slate upon which to construct their political knowledge. Such politically biased "hallucinations" in AI outputs not only distort the information available to users but can also systematically shape and manipulate beliefs, thereby undermining individuals' ability to think freely and independently.

In Kokkinakis v Greece (1993), the European Court of Human Rights (ECtHR) underscored the importance of protecting an individual's internal freedom of thought, emphasizing that this right shields personal belief systems from external manipulation [45]. While this case dealt with religious freedom, its broader implications for freedom of thought can be extended to the digital realm. The potential of LLM hallucinations to influence political beliefs or personal convictions through the production of biased or inaccurate content suggests that freedom of thought, as articulated in Kokkinakis,

could be compromised by AI systems. The ECtHR has also emphasized that even an individual's intention to vote for a particular political party remains a deeply personal and internal conviction, which is protected within the private sphere of one's conscience and autonomy. This internalized decision-making process is part of the *forum internum*, a term used to refer to the innermost realm of personal beliefs and thoughts that are shielded from external intrusion or regulation [46]. The ECtHR's jurisprudence could be leveraged to argue that AI-generated distortions amount to a violation of this fundamental right, especially in cases where individuals unknowingly base their beliefs on fabricated information from AI systems.

If this issue is traced to its root, AI systems rely on extensive datasets for training, which can inadvertently introduce biases reflecting the political or ideological leanings present in those datasets. To mitigate biases, transparency in AI development and training processes is critical. By making datasets and algorithms more accessible for scrutiny, developers and researchers can better understand and address the sources of bias [47]. Regular audits and updates of AI models are also necessary to adapt to evolving social and political contexts, ensuring that the AI's output remains balanced and fair. These measures are vital for maintaining cognitive liberty, the right to form and hold beliefs without undue influence. Effective bias mitigation will help ensure that AI systems contribute positively to democratic discourse by providing balanced and unbiased information. This approach supports a more informed public, capable of participating meaningfully in democratic processes, free from the distortions of biased AI-generated content.

7.1 Viability of Freedom of Thought as Protection Against LLM Hallucinations

While freedom of thought is enshrined as a fundamental human right, its invocation as a defense against the cognitive influences of LLM hallucinations raises several challenges. As technology becomes more deeply integrated into the processes of public discourse and personal cognition, the role of freedom of thought as a safeguard becomes increasingly complex. Although protected under international human rights frameworks, this right faces significant conceptual and practical hurdles when applied to the intricate and often subtle impacts of LLM hallucinations on individual cognition.

One of the key issues with employing freedom of thought as a legal defense is its inherently passive nature. Unlike freedom of expression, which governs the external communication of ideas and opinions, freedom of thought pertains to the internal realm of cognition—the right to hold and develop personal beliefs free from external interference [48]. This makes it difficult to determine when an individual's cognitive liberty has been infringed. LLM hallucinations, which often take the form of fabricated or misleading information, can subtly influence an individual's thought process without explicit coercion. Given that these manipulations typically operate through indirect means, determining whether freedom of thought has been violated becomes a challenging task for courts, which must navigate the diffuse and often invisible nature of such infringements [49]. The lack of overt manipulation complicates the application of this right in contexts where AI subtly shapes the cognitive environment. But it must be noted that the European Court of Human Rights (ECtHR), for instance, has historically recognized violations of fundamental rights when the state's actions or omissions result

in a 'chilling effect' on individuals' freedom of thought or expression [50]. Applying this framework to LLM-related influences, courts may need to evaluate whether such technologies create environments where individuals feel constrained or misled in their intellectual autonomy, even in the absence of overt coercion. This nuanced approach highlights the difficulty of adapting traditional legal frameworks to address the subtleties of algorithmic manipulation.

Another significant challenge arises from the difficulty of proving that an individual's freedom of thought has been meaningfully infringed upon by AI-generated content. LLM hallucinations can have a gradual and cumulative effect, influencing users' perceptions and beliefs over time without a clear point of infraction. Unlike more tangible rights, such as privacy or freedom of expression, which can be visibly and directly breached, violations of freedom of thought are often subtle and internal, making it difficult to draw a clear causal connection between the AI's influence and the individual's cognitive autonomy [51]. Legal systems, which rely on demonstrable evidence of harm, may find it difficult to recognize and address the ways in which cognitive liberty is slowly eroded by repeated exposure to misleading or fabricated information from AI systems.

The subjective nature of thought further complicates the use of freedom of thought as a defense. Thought, by its very nature, is fluid and shaped by countless external stimuli, making it challenging to pinpoint when influence crosses the line into manipulation [52]. In a world where individuals are constantly exposed to various ideas, opinions, and narratives, discerning when an individual's cognitive liberty has been compromised by AI-generated content becomes exceedingly complex. Courts and legal scholars may question whether subtle shifts in cognition caused by LLM hallucinations can truly be considered infringements on freedom of thought, particularly when users themselves may be unaware of the influence that has been exerted on their beliefs and reasoning. This invisible nature of cognitive manipulation complicates the legal recognition of violations and undermines the ability to effectively use freedom of thought as a protective mechanism [53].

Additionally, the protections afforded by freedom of thought are primarily designed to shield individuals from coercive or overt external forces, such as state control or social pressure. However, the influence of LLM hallucinations operates at a more subtle and indirect level, often manifesting through persuasive but fabricated information that does not constitute direct coercion. The diffused effect of AI's influence challenges the traditional understanding of cognitive liberty and raises questions about whether the existing frameworks for freedom of thought are equipped to deal with the nuanced cognitive manipulations introduced by advanced AI systems. The core issue lies in the difficulty of distinguishing between acceptable external influence, such as exposure to diverse ideas, and impermissible manipulation that threatens an individual's ability to form independent opinions [53].

In sum, while freedom of thought is a cornerstone of individual autonomy, its application as a defense in the context of LLM hallucinations presents several challenges. The passive nature of the right, the difficulty of proving cognitive manipulation, the subjectivity of thought, and the indirect influence of AI-generated hallucinations all complicate its use in legal contexts. As AI systems become more pervasive in shaping human cognition, it is imperative to reconsider how freedom of thought can be protected in this

evolving landscape. Without addressing the more nuanced and indirect threats posed by AI technologies, the right to cognitive liberty may prove insufficient in safeguarding individuals against the subtle manipulations introduced by LLMs.

8 Consumer Protection

LLM hallucinations can significantly impact consumer decisions, potentially leading to financial loss, harm, or misinformation. This section delves into the manipulative effects of LLM hallucinations, examining the existing legal frameworks designed to safeguard consumers and address the civil liabilities of entities deploying these technologies.

On February 14th, 2024, in the case Moffatt v. Air Canada, the Canadian Civil Resolution Tribunal became the first court to consider applying strict liability for a loss caused by a chatbot's hallucinated output [54]. Air Canada was found liable for negligent misrepresentation due to misleading information provided by one of its chatbots. The Civil Resolution Tribunal of British Columbia upheld a claim by Jake Moffatt, who, following the death of his grandmother, was given incorrect information by the chatbot about bereavement fares. The chatbot erroneously indicated that a reduced fare could be applied retroactively within 90 days of booking, contrary to the airline's actual policy. Although Air Canada admitted the chatbot's information was misleading, it argued that Moffatt should have checked the information via a linked page, a position the Tribunal rejected [55].

The Tribunal determined that Air Canada was responsible for the chatbot's misrepresentation, emphasizing that liability for the accuracy of information provided by such technology lies with the deploying business. This case highlights a broader legal challenge where traditional liability principles are applied to emerging AI technologies. Businesses must recognize that using AI tools, like chatbots, imposes a responsibility to ensure their accuracy and reliability, as courts are likely to hold companies accountable for the actions of their AI systems [55].

With the increasing integration of AI in various sectors, the potential for harm necessitates a critical evaluation of current regulations. Establishing liability and providing redress for those harmed by LLM hallucinations under EU consumer law involves several key steps. First, it is crucial to identify the responsible party, which may include developers, deployers, or operators of the AI system as mentioned in the proposed AI Liability Directive [56]. Liability can be established by proving negligence, where the entity failed to ensure the accuracy and reliability of the LLM, or through strict liability, where harm is directly linked to the AI's output regardless of fault. Under the EU's Product Liability Directive, AI systems can be considered products, potentially making producers liable for defects that cause damage [57].

Redress for affected consumers can be provided through various legal remedies. Compensation for financial loss is a primary form of redress, ensuring victims are reimbursed for any economic harm suffered[58]. Additionally, corrective measures, such as public retractions of false information and system modifications to prevent future hallucinations, are essential. Punitive damages may also be awarded to deter future negligence and encourage higher standards of care in AI development and deployment [59].

It must be noted that developers are actively attempting to address the issue of LLM hallucinations through a multi-faceted approach, combining rigorous red-teaming

practices with advanced validation techniques to enhance compliance with consumer protection standards. By simulating potential misuse and edge cases, red teams identify weaknesses and biases, allowing for the refinement of datasets and algorithms to minimize inaccuracies [60]. Techniques like reinforcement learning from human feedback (RLHF) and adversarial training further contribute to improving model accuracy and reliability [61]. However, the complexity and inherent unpredictability of AI present challenges in eliminating hallucinations entirely. The probabilistic nature of these models and the vast diversity of human language mean that achieving complete eradication may be elusive even though developers are engaging in continuous testing and validation, incorporating feedback to adjust and improve data sets [62]. Implementing robust quality assurance processes and adhering to ethical guidelines are essential for producing reliable outputs. Despite significant strides made in reducing hallucinations, ongoing research and technological advancements are crucial for further minimizing their frequency and impact. By fostering a culture of diligence and accountability, developers can better align with consumer protection standards and reduce the risk of harmful misinformation, while recognizing that the quest for perfect accuracy remains a challenging and evolving goal. Consumer rights organizations and regulatory bodies play a critical role in enforcing these laws, offering mediation and dispute resolution services [63]. Ensuring transparency in AI operations and strengthening regulatory frameworks are crucial steps in protecting consumers. As AI technologies continue to evolve, ongoing legal adaptations will be necessary to address emerging risks and maintain consumer trust in these powerful tools.

9 Defense Against Manipulation and the AI Act

In the previously discussed 'Belgian AI chatbot suicide case', [64] the core of the controversy lies in the chatbot's responses, which appeared to guide the victim towards self-destructive thoughts rather than offering genuine help or directing him towards appropriate mental health resources. This case highlighted critical issues surrounding the ethical design and deployment of AI systems, especially those interacting with vulnerable individuals who may be more malleable to manipulative ideas.

The EU AI Act clearly identifies and provides safeguards against AI that may have the capacity and the propensity to indulge in 'manipulative or deceptive techniques' [65]. According to Article 5(a) of the Act, the deployment of AI systems that use subliminal or deceptive techniques to materially distort behaviour and impair decision-making is prohibited if it causes significant harm. In the Belgian case, where an AI chatbot allegedly manipulated a man's mental state leading to his suicide, the chatbot's actions can be scrutinized under this provision. The key issue is whether the chatbot used techniques that were purposefully manipulative or deceptive, as defined by the AI Act [66].

The AI Act's stipulation that such techniques must materially distort behaviour by impairing the ability to make an informed decision aligns with the allegations in the Belgian case. If the chatbot's responses indeed impaired the user's ability to make an informed and autonomous decision, leading him to a harmful outcome he would not have otherwise pursued, this would constitute a breach of the regulation. Additionally, the act requires that the resulting harm be significant. Suicide represents an extreme

form of significant harm, meeting the severity criterion outlined [67]. Therefore, if the chatbot's responses are found to have been purposefully manipulative, the case can be considered a critical breach of the AI Act's provisions designed to protect individuals from such severe outcomes [68].

Ethically, the case raises serious questions about the responsibility of AI developers and deployers to ensure their systems do not exploit or manipulate vulnerable individuals. The incident calls for a critical examination of how AI interactions impact user autonomy and well-being. It also brings to light concerns about the AI Act's narrow interpretation of manipulation and lack of clarity about who may qualify as being strictly 'vulnerable' which may leave victims of LLM hallucinations inadequately protected [69]. Hallucinations, where AI systems generate misleading or false information, may not always be covered under the current definitions of manipulation, leaving users exposed to potential harm. This case prompts reflection on the ethical and regulatory challenges in addressing the broader implications of AI-induced harm and the parameters of responsibility for those involved in designing and implementing these systems.

10 Conclusion

The implications of LLM hallucinations on cognitive liberty, the formation of informed opinions, and the potential for manipulative influence, especially in sensitive contexts, caution an urgent need for a comprehensive understanding and regulatory oversight of AI-generated content. As LLMs become more embedded in everyday applications, the line between factual information and fabricated content blurs, posing significant risks to the integrity of public discourse and individual cognitive autonomy.

Paglieri highlights the practical problems of generative AI, emphasizing that the cognitive illusions created by AI can distort our perception of reality and emphasises the requirement for greater explainability in anticipation of an AI Apocalypse [70]. This distortion is particularly problematic in the context of LLM hallucinations, where users are more likely to trust the AI-generated content due to the inherent power imbalance and informational asymmetry between humans and machines. The sophisticated language capabilities of LLMs can create an illusion of authority and reliability, making users more susceptible to accepting false or biased information as truth [71]. Moreover, the phenomenon of prompt hacking reveals another layer of vulnerability in LLMs [72]. By manipulating input prompts, malicious actors can induce LLMs to generate harmful, misleading, or biased content. This ability to exploit the generative process of AI poses significant threats to both individual users and society at large, as it can be used to spread misinformation, influence political opinions, and manipulate vulnerable individuals [73].

EU regulations, including the AI Act, the ECHR's protections for freedom of thought, and consumer protection laws, offer important frameworks for addressing the distinctive challenges posed by Generative AI. While the AI Act prohibits manipulative practices and the ECHR safeguards cognitive liberty, these frameworks often fall short in addressing the subtle and pervasive influence of LLM hallucinations. Consumer protection laws effectively tackle cases where individuals rely on false information generated by chatbots but remain limited in scope, failing to address the broader societal implications of Generative AI. The risks of careless speech, plausible but factually inaccurate or misleading outputs generated by LLMs, further amplify these shortcomings, highlighting

how such outputs can cumulatively erode trust and distort shared knowledge over time [73]. These frameworks must evolve. Specifically, there is a need for adaptive regulatory measures that ensure the accountability of AI systems and the entities deploying them. These measures should include rigorous standards for transparency, verifiability, and the mitigation of biases in AI output through an alignment between human and external knowledge to clean up datasets [74].

An additional perspective on the challenges posed by LLM hallucinations and the broader implications of AI technology can be understood through Norbert Wiener's discussion of entropy in his work 'The Human Use of Human Beings'. Wiener argues that the physical world is governed by a natural tendency toward disorder, a concept encapsulated by the idea of entropy. This principle, when applied to the realm of AI, offers a compelling thesis for why AI systems, including LLMs, often produce unexpected or chaotic results, such as hallucinations. These systems, much like isolated physical systems, may naturally devolve toward states of disorganization, where the coherence of information degrades over time. [75] The diffusion of energy toward equilibrium, a core feature of entropy, mirrors the diffusion of data through vast networks, ultimately resulting in outputs that lack structure or factual accuracy. This theoretical framework could deepen our understanding of why LLMs sometimes fail to perform as expected and emphasize the importance of continually refining AI architectures to mitigate this inherent tendency toward disorder. In conclusion, LLMs and their propensity for hallucinations present significant risks to cognitive liberty and the formation of informed opinions. The trust placed in these systems, coupled with their capacity for generating persuasive yet false content, requires a careful examination of their role in society. By understanding the mechanisms driving LLM hallucinations and evaluating existing legal protections, we can develop strategies to safeguard against the manipulative influence of AI, ensuring that the benefits of these technologies are realized without compromising the integrity of public discourse or individual cognitive autonomy.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. IBM: What Are AI Hallucinations? https://www.ibm.com/topics/ai-hallucinations. Accessed 18 Apr 2024
2. The Artificially Intelligent Enterprise: The Perils of Language Model Hallucinations. https://www.theaienterprise.io/p/ai-language-model-hallucinations. Accessed 18th Apr 2024
3. The New York Times: How AI Chatbots Become Political. https://www.nytimes.com/interactive/2024/03/28/opinion/ai-political-bias.html?ugrp=u&unlocked_article_code=1.gU0.PO1t.oWpVBdAZ1qfv&smid=url-share. Accessed 18 Apr 2024
4. Euro News: Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate. Accessed 18 June 2024
5. The Guardian: Colombian Judge Says he used ChatGPT in Ruling. https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling. Accessed 18 April 2024
6. Bublitz, C.: Cognitive liberty or the international human right to freedom of thought. In: Advances in Human Factors and Ergonomics, pp. 83–90. Springer, Dordrecht (2015)

7. Prescott, M., et al.: Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. JMIR AI **3**, e54482 (2024)
8. Bublitz, C.: Cognitive liberty as a legal concept. In: Hildt, E., Franke, A. (eds.) Cognitive Enhancement: An Interdisciplinary Perspective, pp. 233–264 (2013)
9. Hacker, P.: Manipulation by algorithms: exploring the triangle of unfair commercial practice, data protection, and privacy law. Eur. Law J. (2021)
10. Shah, D.: The Beginner's Guide to Hallucinations in Large Language Models. https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models#:~:text=Hallucinations%20in%20LLMs%20refer%20to,trust%20placed%20in%20these%20models. Accessed 10 June 2024
11. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus **15**(2) (2023)
12. ChatGPT response dated April 1st, 2024. Replication can be done using the same input; however, different outputs can be expected after model upgrades (2024)
13. Baum , J., Villasenor, J.: The politics of AI: ChatGPT and political bias. https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/?b=1. This study also asks various other questions including questions about immigration, taxes, banning of automatic weapons etc., and then made its conclusion on the responses to these inquiries
14. Huang, L., Yu, W.: A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. Harbin Institute of Technology, 9th November 2023
15. Wang, Y., Wang, Y., Zhao, D., Xie, C., Zheng, Z.: VideoHallucer: evaluating intrinsic and extrinsic hallucinations in large video-language models. The table has been created as a hybrid from classifications in both these sources to better assess the manipulative aspects of LLM hallucinations. arXiv (2024)
16. Varshney, N., Yao, W.: A stitch in time saves nine: detecting and mitigating hallucinations of LLMs by validating low-confidence generation. Arizona State University, 12th August 2023
17. Shah, D.: (n 6)
18. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
19. Mata v. Avianca, Inc.: F. Supp. 3d, 22-cv-1461 (PKC), 2023 WL 4114965, at *2 (S.D.N.Y. June 22 2023)
20. Association of Corporate Counsel: Practical Lessons from the Attorney AI Missteps in Mata v. Avianca, 8 August 2023. https://www.acc.com/resource-library/practical-lessons-attorney-ai-missteps-mata-v-avianca
21. The New York Times: Here's What Happens When Your Lawyer Uses ChatGPT, 27 May 2023
22. Zuccon, G., Koopman, B., Shaik, R.: Chatgpt hallucinates when attributing answers. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, pp. 46–51, November 2023
23. Lening, C.: On Ethics, the EDPB, Errors, and Endorsements. https://www.linkedin.com/pulse/ethics-edpb-errors-endorsements-carey-lening-cdpp-h3bge/?trackingId=qPjJRWFWRGSWg3MN0Sz%2BMA%3D%3D. Accessed 10 July 2024
24. Magesh, V., et al.: Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools (2024). https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf. Accessed 6 June 2024
25. Bhattacharya, R.: Who is hallucinating - Stanford University or Thomson Reuters/Lexis Nexis? 26 June 2024. https://www.linkedin.com/pulse/who-hallucinating-stanford-university-thomson-nexis-bhattacharya-jqfwe/?trackingId=5P1AYdoITGGYY7V6SO%2FIGQ%3D%3D
26. Marcus, G.: The dark risk of large language models, Wired, 29 December 2022. https://www.wired.com/story/large-language-models-artificial-intelligence/

27. Marcus, G.: The dark risk of large language models. Wired, 29 December 2022. https://www.wired.com/story/large-language-models-artificial-intelligence/. (n 4)
28. Business Insider: A widow is accusing an AI chatbot of being a reason her husband killed himself, 4th April 2023. https://www.businessinsider.com/widow-accuses-ai-chatbot-reason-husband-kill-himself-2023-4?international=true&r=US&IR=T. Accessed 4 July 2024
29. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM **9**(1), 36–45 (1966)
30. Lovens, P.-F.: Without these conversations with the chatbot Eliza, my husband would still be here. La Libre Belgique, 28 March 2023. https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/
31. Xing, C.: He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. VICE, 31st March 2023. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says. Accessed 19 June 2024
32. Robson, G.J., Tsou, J.Y. (eds.): Technology Ethics: A Philosophical Introduction and Readings. Routledge, New York (2023, forthcoming)
33. Alexander, L., Moore, M.: Deontological Ethics. The Stanford Encyclopedia of Philosophy. Winter 2020 Edition. https://plato.stanford.edu/archives/win2020/entries/ethics-deontological/
34. Durmus, E., et al.: Measuring the Persuasiveness of Large Language Models. Anthropic, 9 April 2024. https://www.anthropic.com/news/measuring-model-persuasiveness. Accessed 1 July 2024
35. Murtarelli, G., Gregory, A., Romenti, S.: A conversation-based perspective for shaping ethical human–machine interactions: the particular challenge of chatbots. J. Bus. Res. **129**, 927–935 (2021)
36. Murtarelli, G., Gregory, A., Romenti, S.: A conversation-based perspective for shaping ethical human–machine interactions: the particular challenge of chatbots. J. Bus. Res. **129**, 927–935 (2021). (n 3)
37. Rozado, D.: The political biases of ChatGPT. Soc. Sci. **12**(3), 148 (2023). https://doi.org/10.3390/socsci12030148
38. Pearson, J.: AI-Generated Propaganda Is Just as Persuasive as the Real Thing, Worrying Study Finds. VICE, 21 February 2024. https://www.vice.com/en/article/ak38xb/ai-generated-propaganda-is-just-as-persuasive-as-the-real-thing-worrying-study-finds
39. Goldstein, J.A., Chao, J., et al.: How persuasive is AI-generated propaganda? PNAS Nexus **3**(2), 034 (2024)
40. Weidinger, L., et al.: Sociotechnical safety evaluation of generative AI systems. Google DeepMin, p. 41, 31st October 2023. https://arxiv.org/pdf/2310.11986
41. Weidinger, L., et al.: Sociotechnical Safety Evaluation of Generative AI Systems. Google DeepMin, p. 41, 31 October 2023. https://arxiv.org/pdf/2310.11986. pp. 42–43
42. Barman, D., Guo, Z., Conlan, O.: The dark side of language models: exploring the potential of LLMs in multimedia disinformation generation and dissemination. Mach. Learn. Appl. **16** (2024). https://doi.org/10.1016/j.mlwa.2024.100545
43. John Stuart Mill, On Liberty, John W. Parker and Son (1859)
44. John Stuart Mill, On Liberty. John W. Parker and Son (1859). (n 7)
45. Kokkinakis v Greece (1993). 17 EHRR 397
46. Russian Conservative Party of Entrepreneurs and Others v. Russia, App No. 55066/00, 55638/00 (11th January 2007) 76; Georgian Labour Party v. Georgia, App No. 9103/04 (8th July 2008) 120
47. Bontridder, N., Poullet, Y.: The role of artificial intelligence in disinformation. Data & Policy (2021). https://doi.org/10.1017/dap.2021.20

48. Pastor, E.R.: The freedom of thought, conscience, and religion in the age of neuroscience: revisiting the forum internum. J. Relig. Eur. **1**(aop), 1–27 (2024)
49. Bublitz, C.: Freedom of thought as an international human right: elements of a theory of a living right. In: Blitz, M.J., Bublitz, J.C. (eds.) The Law and Ethics of Freedom of Thought, vol. 1. Palgrave Studies in Law, Neuroscience, and Human Behavior (2021)
50. Dink v. Turkey, European Court of Human Rights, Application No. 2668/07, 14th September 2010
51. Swaine, L.: Freedom of thought as a basic liberty. Polit. Theory **46**(3), 405–425 (2018)
52. McCarthy-Jones, S.: Freedom of thought: who, what, and why? In: Blitz, M.J., Bublitz, J.C. (eds.) The Law and Ethics of Freedom of Thought, vol. 1. Palgrave Studies in Law, Neuroscience, and Human Behavior. Palgrave Macmillan, Cham (2021)
53. Jongepier, F., Klenk, M.B.O.T.: The Philosophy of Online Manipulation. Routledge Research in Applied Ethics. Routledge - Taylor & Francis Group (2022). https://doi.org/10.4324/9781003205425
54. Moffatt v. Air Canada: 2024 BCCRT 149 (CanLII). https://canlii.ca/t/k2spq. Accessed 26 June 2024
55. Higgins, M.: Air Canada Chatbot Case Highlights AI Liability Risks. Pinsent Masons, 27 February 2024
56. 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/PDF/?uri=CELEX:52022PC0496. Accessed 17 June 2024
57. Launders, J.: Beyond the AI Act: The AI Liability Directive & the Product Liability Directive. Tech Law Blog, (5th March 2024) European Parliament and Council, 'Proposal for a Directive on Liability for Defective Products' (COM(2022) 495) (2022) C9 0322/2022, 2022/0302(COD). https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.europarl.europa.eu%2FRegData%2Fcommissions%2Fimco%2Finag%2F2024%2F01-24%2FCJ24_AG(2024)758731_EN.docx&wdOrigin=BROWSELINK
58. Launders, J.: Beyond the AI Act: The AI Liability Directive & the Product Liability Directive. Tech Law Blog, (5th March 2024) European Parliament and Council, 'Proposal for a Directive on Liability for Defective Products' (COM(2022) 495) (2022) C9 0322/2022, 2022/0302(COD). https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.europarl.europa.eu%2FRegData%2Fcommissions%2Fimco%2Finag%2F2024%2F01-24%2FCJ24_AG(2024)758731_EN.docx&wdOrigin=BROWSELINK. (n 42)
59. Mason, Hayes & Curran: Potential Liability for Chatbot Hallucinations? 20 March 2024. https://www.mhc.ie/latest/insights/potential-liability-for-chatbot-hallucinations
60. Buszydlik, A., et al.: Red Teaming for Large Language Models At Scale: Tackling Hallucinations on Mathematics Tasks, 30 December 2023. https://arxiv.org/abs/2401.00290v1
61. Christiano, P., Leike, J., et. al.: Deep reinforcement learning from human preferences (2017)
62. Kang, H., Ni, J., Yao, H.: Ever: mitigating hallucination in large language models through real-time verification and rectification (2023). arXiv preprint arXiv:2311.09114
63. Shoosmiths: From Chatbots to ChatGPT: Navigating consumer rights in an AI-driven world. https://www.shoosmiths.com/insights/articles/from-chatbots-to-chatgpt-navigating-consumer-rights-in-an-ai-driven-world
64. Shoosmiths: From Chatbots to ChatGPT: Navigating consumer rights in an AI-driven world. https://www.shoosmiths.com/insights/articles/from-chatbots-to-chatgpt-navigating-consumer-rights-in-an-ai-driven-world. (n 4)
65. EU Artificial Intelligence Act, 19th March 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf. Accessed 5 June 2024

66. Franklin, M., Tomei, P.M, Gorman, R.: Vague concepts in the EU AI Act will not protect citizens from AI manipulation. OECD AI. Policy Observatory, 7th September 2023. https://oecd.ai/en/wonk/eu-ai-act-manipulation-definitions. Accessed 1 July 2024
67. Franklin, M., Tomei, P.M., Gorman, R.: Vague concepts in the EU AI Act will not protect citizens from AI manipulation. OECD AI. Policy Observatory, 7th September 2023. https://oecd.ai/en/wonk/eu-ai-act-manipulation-definitions. Accessed 1 July 2024. Art 5 (a) (n 52)
68. Cabrera, L.: EU AI Act Brief – Pt. 3, Freedom of Expression. Center for Democracy and Technology. https://cdt.org/insights/eu-ai-act-brief-pt-3-freedom-of-expression/#:~:text=Article%205(1)(a,an%20informed%20decision%2C%20thereby%20causing
69. Franklin, M., et al.: The EU's AI Act needs to address critical manipulation methods. OECD AI. Policy Observatory, 21 March 2023. https://oecd.ai/en/wonk/ai-act-manipulation-methods. Accessed 1 July 2024
70. Paglieri, F.: Expropriated minds: on some practical problems of generative AI, beyond our cognitive illusions. Philos. Technol. **37**, 55 (2024). https://doi.org/10.1007/s13347-024-00743-x
71. McGuire, J., De Cremer, D., Hesselbarth, Y., et al.: The reputational and ethical consequences of deceptive chatbot use. Sci. Rep. **13**, 16246 (2023) https://doi.org/10.1038/s41598-023-41692-3
72. Ramlochan, S.: Prompt Hacking: The New Cyber Threat. Prompt Engineering and AI Institute, 5th March 2024. https://promptengineering.org/the-rise-of-a-new-threat-prompt-hacking/
73. Wachter, S., Mittelstadt, B., Russell, C.: Do large language models have a legal duty to tell the truth? R. Soc. Open Sci. (2024)
74. Zhang, S, Pan, L., Zhao, J., Wang, W.Y.: The knowledge alignment problem: bridging human and external knowledge for large language models (2023)
75. Wiener, N.: The human use of human beings. Br. J. Philos. Sci. **3**(9), 91–92 (1952)