Panagiota Kiortsi, Ioannis Iglezakis, Christos Feidas, Dimitrios Kosmopoulos, George Thomopoulos

Keywords: Large Language Models (LLMs); Hallucinations; AI Regulation; Ethics and Accountability; Comparative Law (EU–US)

Abstract

Large Language Models (LLMs) have rapidly transformed interactions with artificial intelligence, enabling applications in law, healthcare, finance, and education. Yet, they display a critical flaw: the production of false, misleading, or fabricated outputs, commonly referred to as "hallucinations." These outputs can misinform users, compromise professional integrity, and erode trust in institutions, raising pressing legal and ethical challenges.

This paper examines whether hallucinations can be effectively regulated or whether they remain inherently elusive. It first identifies their root causes—biased or incomplete data, semantic misinterpretations, and adversarial attacks—and the risks they pose in high-stakes contexts. It then compares regulatory approaches in the European Union and the United States. The EU, through the AI Act and the GDPR, adopts a rights-centered, precautionary framework based on transparency, accountability, and risk classification, but stops short of directly addressing hallucinations. The U.S., in contrast, relies on a fragmented mix of state laws, sectoral rules, professional standards, and voluntary guidelines, prioritizing innovation but leaving regulatory gaps.

Despite their differences, both systems converge on risk-based oversight, transparency, and impact assessments. However, hallucinations are treated only indirectly, leaving systemic harms unresolved. The paper argues that effective governance requires a hybrid approach that combines technical safeguards (diverse datasets, reinforcement learning, hallucination-detection tools), legal obligations (risk management, accountability, user protection), and ethical oversight. Ultimately, hallucinations must be seen not merely as programming errors but as sustainability challenges for AI governance: balancing innovation with fairness, human rights, and democratic values.

Patterns and root causes of hallucinations

Specifically, hallucinations in Large Language Models (LLMs) manifest either as factual events containing incorrect information in the form of inaccuracies about real-world facts, or as entirely fabricated data that deviate from the content used to train the

model [1] . In the second category, inaccuracies appear as content that includes contradictory statements within the same response, as well as answers that are unrelated to the question due to misinterpretation. The persuasive nature of hallucinated responses makes detection and correction difficult. It also creates a fraudulent guise of AI omnipotence that gains ground form fact-checking when using these tools.[2]

If the training data of LLMs contain errors, biases, or incomplete information, the model may generate flawed or unintelligible outputs (hallucinations). LLMs operate on probabilistic patterns (for example the phrase “Greece is a Member State of” will most probably end with “E.U”). However, when the probabilistic pattern is not correct, the system itself is not always able to recognize this fault. Consequently, because an LLM relies on its training data, it may generate answers ranging from plausible to entirely erroneous because overgeneralization leads to fabricated responses[3]. If the training data are incomplete, incorrect, or incoherent, the system’s ability to self-verify is limited, making it difficult to detect and correct such errors on its own[4].

In some cases, training data can be maliciously crafted and manipulated. This kind of manipulation is known as Data poisoning, which is a type of cyberattack. Data poisoning has become an increased threat to the security and trustworthiness of large language models (LLMs), as maliciously designed data can tamper the LLMs training process, targeting performance degradation, backdoor insertion, and manipulation of outputs.

LLMs may also adapt to misleading patterns detected during training, reproducing answers that lack substantiation but seem fluent and coherent. Another root cause of hallucinations is that LLMs comprehend linguistic context without grasping subtle semantic nuances.[5] Yet such semantic precision is fundamental in certain scientific fields. For instance, in Law, a single term within legislation may carry a particular

---

[1] Sun, Y., Sheng, D., Zhou, Z. et al. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun 11, 1278 (2024). https://doi.org/10.1057/s41599-024-03811-x

[2] Cossio M, A comprehensive taxonomy of hallucinations in Large Language Models, arxiv Cornell University , 3 Aug 2025, https://doi.org/10.48550/arXiv.2508.01781

[4] Cleti, Meade & Jano, Pete. (2024). Hallucinations in LLMs: Types, Causes, and Approaches for Enhanced Reliability. 10.13140/RG.2.2.12184.61445.

[5] Huang l et al Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Journals ACM Transactions on Information Systems Vol. 43, https://doi.org/10.1145/3703155

meaning that an LLM cannot perceive. An illustrative example was observed during the training of the AILA model: the system interpreted the word “right” in Articles 12–23 of the General Data Protection Regulation (GDPR) as the broader legal concept of a right, rather than its specific meaning and function within the GDPR.

Hallucinations in legal cases have become a growing universal concern. Because of the gravity of the consequences on fundamental human rights, hallucinations in LLMs used in judicial procedures, have been systematically documented since 2023 across different countries and legal systems, including the United States, Canada, Australia, and the United Kingdom[6]. According to the findings they often appear as references to non-existent case law, inaccurate interpretations of legislation and erroneous or nonexistent citations. According to a study, legal hallucinations in LLMs increase with complexity. Simple tasks, like confirming case law existence show lower error possibilities, while complex ones—like interpreting legal doctrines—can trigger hallucinations. It has been documented that LLMs hallucinations appear less in prominent Supreme Court cases than on less visible lower courts case law[7].

REGULATION OF LLMS HALLUCINATION IN EU AND USA

The question if LLMs hallucinations should be regulated is not simplicial. If LLMs are propagated globally for high-impact decisions in diverse fields such as justice and health, concerns have been raised about the effects of a regulation that addresses the risks of LLMs hallucinations by creating an “algorithmic monoculture”[8], in which algorithm convergence will dictate decision making. According to research, attempts to

---

[6] AI Hallucination cases διαθέσιμο στο διαδικτυακό χώρο https://www.damiencharlotin.com/hallucinations/

[7] Dahl M et all, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, Journal of Legal Analysis 16, no. 1 (2024): 64-93, https://doi.org/10.1093/jla/laae003

[8] Bommasani R et al, Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?, , arxiv Cornell University, https://doi.org/10.48550/arXiv.2211.13972

develop universal algorithms may produce biased and discriminatory LLMs, making them more prone to error and creating an algorithmic dictatorship.

Moreover, the potential danger "of outcome homogenization" arises, where groups of individuals face discrimination as they receive negative outcomes across multiple interconnected decision-makers because of algorithm convergence. This situation could create serious harms in vital areas of life, including employment, insurance and education, by limiting access to essential opportunities and by creating algorithmic outcasts[9].

In EU, the Artificial Intelligence Act - Regulation (EU) 2024/1689 was the first regulatory initiative that focused on stipulating harmonized rules on AI and establish a human-centric legal approach. Regulation 2024/1689 distinguish between general-purpose artificial intelligence models and general-purpose artificial intelligence models with systemic risks. Providers of general-purpose AI models with systemic risk are subjected to specific obligations[10].

The AI Act establishes a regulatory framework for AI systems, particularly focusing on high-risk applications. Chapter III outlines the classification rules and regulatory requirements for high-risk AI systems. According to Article 6(1), an AI system is deemed high-risk if (a) it is a safety component of a product or a product itself as defined under EU harmonisation laws listed in Annex I, and (b) it is subject to third-party conformity assessments before being placed on the market or put into service.

Section 2 of Chapter III sets out obligations for high-risk systems, reflecting the AI Act's human-centric design. Article 14 introduces the requirement for human oversight, mandating that high-risk AI systems be designed to allow effective human control. Oversight must aim to prevent or minimize risks to health, safety, and fundamental rights—especially where risks remain even after technical safeguards.

The Artificial Intelligence Act (Regulation (EU) 2024/1689) introduces a risk-based classification of AI systems, with "high-risk" systems subject to strict requirements. Several provisions are particularly relevant to hallucinations: Article 10 requires

---

[9] Bommasani R et al

[10] Fabiano AI Act and Large Language Models (LLMs): When critical issues and privacy impact require human and ethical oversight, available at https://arxiv.org/html/2404.00600v2

developers to ensure the quality of data used in training, recognizing that biased or inaccurate datasets increase the likelihood of hallucinations. Article 13 mandates transparency and verifiability, obliging developers to ensure that AI systems can explain their outputs and be subject to human oversight. Articles 14, 15 and 17 insist on effective human supervision, especially in sensitive domains, to mitigate the consequences of flawed outputs to guarantee that they are resilient regarding errors, faults or inconsistencies and manage quality through quality management systems. Article 25 introduces deployers obligation for conducting Fundamental rights impact assessment for high-risk AI systems on the specific risks of harm on the categories of natural persons or groups of persons. Furthermore articles 59-60 stipulate mechanisms for ongoing testing monitoring and auditing, acknowledging that AI systems evolve and may degrade over time, thus generating unexpected or erroneous results.

Where LLMs are categorized as general-purpose artificial intelligence models, according to article 53 providers are obligated to draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, which, according to Annex XI section 1 includes “the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies (e.g. cleaning, filtering, etc.), the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases, where applicable. However, even if obligating providers to use sufficient and concrete training data mitigates some root causes of hallucinations it doesn’t cure all of them (for example semantic nuances). This is why according to some opinions, hallucinations generated by LLMs that are categorized as general-purpose artificial intelligence models are not directly or sufficiently regulated by the Artificial Intelligence Act - Regulation (EU). Moreover, even though the AI Act does not categorize LLMs as high-risk, their implementation can generate high-risk—especially in professional areas and fundamental rights. Recently, hackers bypassed ChatGPT's safety protocols and tricked the AI into providing instructions for

manufacturing a homemade bomb and efficient methods of smuggling cocaine into Europe[11].

The provisions of AI Act have also been scrutinized as insufficient to prevent and mitigate the impact of hallucinations. Specifically, according to this approach, the Act focuses primarily on ex ante obligations (design, testing, documentation), leaving gaps in ex post accountability and remedy. In particular, according to the EU Commission Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act)[12], providers of general-purpose AI models placed on the market before 2 August 2025 are not required to undergo retraining or unlearning where this cannot be carried out for past actions, where parts of the training data are unavailable, or where retrieval would impose a disproportionate burden on the provider. In such cases, the provider must clearly disclose and justify the limitation in both the copyright policy and the summary of the content used for training. In such cases the possibility to override to some extend provider's obligations of openness may lead to potential harm on fundamental rights and has given rise to significant concerns in relation to the likelihood of such harm or adverse impact.

However, where AI Act provide limited support to fundamental human rights jeopardized by LLMs generated errors, the GDPR can make up the deficit by adopting a rights-based, accountable legal framework. GDPR does not regulate LLMs hallucinations directly. It does however provide an ex-ante and ex post protective environment.

Under GDPR if an activity includes processing of personal data that is "likely to result in a high risk to the rights and freedoms of natural persons" according to article 35 of GDPR, a data protection impact assessment must be contacted by the data controller (the provider of LLM). There is a significant difference between the approach of risk for human rights AI Act and GDPR. Specifically, according to GDPR, data processing that led to high risk to the rights and freedoms of natural persons is not acceptable. AI Act does not prohibit high risk AI systems.

---

[11] Sharma A, Man Asks ChatGPT How To Smuggle Drugs Into Europe, Bot Lists Suggestions, Feb 06, 2023, available at https://www.ndtv.com/feature/man-asks-chatgpt-how-to-smuggle-drugs-into-europe-bot-lists-suggestions-3756323

[12] Communication to the CommissionApproval of the content of the draft Communication from the Commission – Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act)

Overall, a DPIA is an essential protective mechanism whenever LLMs are produced and deployed in ways that could significantly affect individuals, even if the LLM is not itself classified as high-risk by default. According to the Article 29 Working Party guidelines, DPIAs should identify for the sources of risk (Recital 90) and the likelihood and severity of the risk (Recital 90), the potential consequences on the rights and freedoms of data subjects.

Moreover Article 22 of the GDPR, stipulates that individuals have the right not to be subject to decisions based solely on automated processing that significantly affect them. This provision highlights the obligation of transparency, explainability, and the use of accurate data in algorithmic decision-making. Article 22 provides Algorithmic accountability that represents the obligation to explain to the user how an AI system functions what data are processed and the parameters of the algorithm.

Hallucinations are generated when AI-generated automated decisions and outputs are based on inaccurate and incomplete or biased data that violate principles of lawful, fair transparent data that are accurate and up to date regarding the purposes for which they are processed according to article 5 of GDPR. Fairness and accuracy are key factors in mitigating hallucinations that are evoked by the generation of non-existent, irrelevant, biased and erroneous data.

However, the protective environment of GDPR is not panacea. Even though the Hamburg Commissioner for Data Protection and Freedom of Information (HmbBfDI) conclude that the GDPR applies to Large Language Models (LLMs) under the under the prerequisite that personal data are processed. If processed training data are anonymized or not identifiable data, the mere storage of an LLM does not constitute processing under the GDPR. Furthermore, this approach does not conclude that when the life circle of LLM training and the outcomes involve personal data, that data subjects should be stripped of their data subjects' rights or providers and deployers as data controllers don't have the obligation to comply with GDPR[13].

Other Data Protection Authorities have a different approach (UK Information Commissioner's Office (ICO)) as they focus on the purpose of the AI system. When the

[13] Christakis T, AI Hallucinations and Data Subject Rights under the GDPR: Regulatory Perspectives and Industry Responses, December 2024). Available at SSRN: https://ssrn.com/abstract=5042191 or http://dx.doi.org/10.2139/ssrn.5042191

output may pose high risks for natural persons the accuracy and contextual transparency of data is required. ICO also emphasizes the need of providing information among deployers and developers in the same approach that AI Act has stipulated in article 25.

However as it has already been explained factual precision and accuracy does not guarantee output accuracy because of the probabilistic patterns the LLMs use by predicting an answer without comprehending the meaning.

While the GDPR and the AI Act provide a robust framework for safeguarding natural persons on the harmful impacts of LLMs hallucinations the algorithmic transparency cannot protect from all potential dangers. According to the EU Commission guidelines transparency and accountability can be eroded in cases LLMs have been used before August 2025. However, this approach does not fucus on harmful outcomes on individuals and hinders users' protection over innovation and does not balance innovation and fundamental rights.

Furthermore, a more dynamic approach could be more effective with continuous risk reassessment that focus on outcomes.

Regulation in USA

In the United States no overarching, federal regulation currently applies on artificial intelligence. Artificial intelligence is regulated by industry self-governance initiatives, State's legislation and court decisions. instead, federal efforts have focused on supporting research and innovation, developing voluntary guidelines, and overseeing government use of AI under existing agency authorities. The U.S. state of Colorado has become the first state to initiate a comprehensive legislation targeted to artificial intelligence. The Colorado Artificial Intelligence Act (CAIA) adopts similar approach to the Artificial Intelligence Act by introducing requirements on transparency, disclosure, and consumer protections for "high-risk" AI systems, imposing compliance obligations as conducting risk-management and impact assessments. There is also obligation to mitigate algorithmic discrimination.

Specifically, the Colorado Artificial Intelligence Act (CAIA) will get into effect February 1, 2026, and focuses on high-risk AI systems that substantially influences "consequential decisions," which are decisions with significant legal or economic effects.

Developers and deployers of high-risk AI systems have the responsibility to prevent algorithmic discrimination, which is defined as unlawful differential treatment based on protected characteristics like race, gender, religion, or disability. When an AI system is used in a consequential decision, consumers must be informed that AI is being used. Developers must provide a description of the system and its purpose. Developers are obliged to publish a summary on how they manage risks of algorithmic discrimination and test their systems for bias. Deployers must provide consumers with information about the AI system. They also have the obligation to provide an explanation for adverse decisions. Deployers must provide a publicly available statement summarizing the types of high-risk systems, the management and mitigation "of any known or reasonably foreseeable risks of algorithmic discrimination that may arise from deployment of high-risk systems, and the nature, source, and extent of the information collected and used by the deployer"[14]

The Act establishes consumer protections. Consumers can request the correction of any incorrect personal data that a high-risk system processed in making a consequential decision. However, it is not clear if this provision is similar the right to rectification stipulated to the GDPR.

Furthermore, consumer can appeal, via human review, an adverse consequential decision concerning the consumer arising from the deployment of a high-risk system.

Deployers must inform the attorney general regarding the discovery of algorithmic discrimination, within 90 days after the discovery, that the high-risk system has caused.

In general, the USA regulation is segmented, by overlapping or conflicting set of legal rules and regulations.

However, Artificial intelligence (AI) in the United States is not yet governed by a broad federal regulatory framework; instead, federal efforts have focused on supporting

[14] SB24-205 Consumer Protections for Artificial Intelligence First Regular Session | 75th General Assembly, Colorado General Assembly available at https://leg.colorado.gov/bills/sb24-205

research and innovation, developing voluntary guidelines, and overseeing government use of AI under existing agency authorities.

The National AI Initiative Act and the CHIPS and Science Act promote technical standards collaborative frameworks and guidelines to mitigate bias in AI training rather than imposing binding regulations. The National AI Initiative Act adopts a risk-mitigation framework for AI systems to promote trustworthy AI systems.

Approaches to regulation vary, to sector-specific measures in finance, balancing innovation with risk management and accountability, and balancing international cooperation and competitiveness.

LLMs hallucinations are not regulated in a targeted and direct manner. General consumer-protection and securities rules against deception apply on LLMs hallucination and so does regulation on health and finance sector legal framework.

Federal consumer-protection & advertising law apply to AI systems that generate deceitful content and misrepresentation of LLMs capabilities or use. Other sectoral regulations addressed chatbots giving wrong or incomplete answers that trigger risk of harm for individuals.

According to the Federal Trade Commission Compliance Plan on advancing governance innovation and risk management for agency use of AI, the FTC is taking a measured, risk-based approach to AI adoption—emphasizing ethical use, transparency, inter-agency collaboration, and workforce readiness—to ensure AI strengthens its mission of protecting consumers and promoting fair competition. Potentially high-risk systems will require risk mitigation plans approved by enterprise risk management boards. The agency's plan emphasizes governance, innovation, workforce development, and risk management to ensure AI use advances the public interest while safeguarding rights, privacy, and security[15].

The Algorithmic Accountability Act of 2023 (U.S.), is a proposed U.S. federal law that focuses regulate on AI and automated decision-making that affect natural persons in critical areas — credit, housing, education, employment, etc. It states that developers

[15] Federal Trade Commission Compliance Plan for OMB Memoranda M-24-10:ON ADVANCING GOVERNANCE, INNOVATION, AND RISK MANAGEMENT FOR AGENCY USE OF ARTIFICIAL INTELLIGENCE September 2024, available at https://www.ftc.gov/system/files/ftc_gov/pdf/FTC-AI-Use-Policy.pdf

must conduct impact assessments of systems (AI or automated) that make "critical decisions" to check for bias and discrimination and safeguard fairness, transparency, privacy, security, and effectiveness. The obligation to conduct impact assessments is mandatory for AI systems already in use and new ones. Companies must report impact assessments to the Federal Trade Commission (FTC). The FTC is delegated to provide guidelines on the assessments. Although the proposed legislation addresses hallucinations in an indirect manner it has not eneterd into force.

Finally, professional codes of conduct have a targeted approach to LLMs hallucinations. In some cases where chatbot hallucinations were presented to courts of Law, the court adjudicated that law practitioners have the obligation to ensure accuracy in their filings and not stand by fake legal research. Although using LLMs on legal research is not illegal, improper employment of generative AI may result in a breach of care, leading to a legal malpractice claim or lawsuit and professional misconduct[16]

Even though here's no specific law that regulates LLMs hallucinations deception, the harmful effects in natural persons safety, discrimination and deception are regulated through preexistent legal frameworks on consumer rights and codes of professional conduct.

It is evident that the **U.S.** vs. the **EU** approach to the regulation of LLMs hallucination mitigation presents some similarities, especially as the existing and proposed regulation does not directly regulates hallucinations. It focuses on imposing obligations to Companies and deployers such as conducting impact assessments and reporting for automated decision systems that make "critical" decisions or have high -risk impacts on natural persons. EU AI Act also requires transparency and imposes some obligations for general-purpose models, as LLMs while some practices are outright prohibited.

Both legal systems adopt a risk assessment-centric approach that mitigates bias, and reinforces safety, transparency, effectiveness and reporting to regulators. However, EU AI Act requires more granular technical and process requirements.

---

[16] Munir, Bakht, Hallucinations in Legal Practice: A Comparative Case Law Analysis (February 24, 2025). International Journal of Law, Ethics, and Technology, Available at SSRN: https://ssrn.com/abstract=5265375 or http://dx.doi.org/10.2139/ssrn.

U.S. legislation applies to covered entities operating in the U.S while the EU AI Act and GDPR are explicitly extraterritorial and apply to providers or deployers placing AI systems on the EU market or whose outputs are used in the EU.

The EU AI Act and GDPR classify systems by risk, imposing a formal risk-management system while US adopts self-regulation in a broader extend for conformity assessments. That "gold standard" approach often satisfies U.S. assessment requirements and speeds compliance with EU rules that apply extraterritorially. However, both legal systems focus less on LLMs and their potential harmful outcomes when outputs are wrong or misleading.

Ethical Considerations

Large Language Models (LLMs) have obtained a fundamental role in education, research, business, and governance. Unfortunately, the unintentional or deliberate generation of erroneous or fabricated content—raises serious ethical concerns that go beyond technical reliability and erode the protective perimeter of human rights.

Hallucinations blur the line between fact and fiction. When false information is presented with omnipotent confidence, users tend to accept it as accurate. In systems that make critical decisions such as law, finance and health, these errors can have cascading social consequences contributing to misinformation and eroding public trust in fundamental institutions and AI[17].

As users tend to rely increasingly on LLMs to assist their decision-making they tend to jeopardize their autonomy. False or fabricated information compromises their ability to make informed decisions in critical areas of their life or their professional conduct. In cases when the user is not equipped with expertise and resources to verify and cross check LLMs outputs they get exposed in dangers especially if they are vulnerable because of their age or inability to track false outputs.

All the above-mentioned issues raise the question of liability when LLMs hallucinate and who should be held responsible when harms occur. The legal frameworks in EU and US impose obligations to the developers and deployers. Nevertheless, in recent cases the users were held accountable for LLMs hallucinations. Ethical and professional duties prohibit the use of LLMs without research fact cross checking. In some cases, the lines that define victims of AI-driven misinformation are blurred. According to the recent case law professionals in law, (and medicine, education, and other fields) have strict ethical duties of competence and truthfulness. The integration of LLMs into professional

[17] Patel H, Bane and boon of hallucinations in context of generative AI, Information Technology, University of Texas San Antonio, Carlos Alvarez college of Business, April 01, 2024, available at:

practice therefore requires ethical vigilance. In other cases where decisions navigates through gray areas, human judgment is irreplicable[18].

Moreover, complex technological supply chains may also blur the roles and responsibilities of developers and deployers. Recently the EU Commission guidelines provide insight on the roles of developers and deployers especially in cases that models may be modified or fine-tuned into new models.

Since LLMs are trained on vast datasets, the danger to reproduce existing societal stereotypes, toxic patterns and imbalances is lingering. Biased, discriminative and hostile narratives can multiply content that attacks individuals based on their race, gender, religion, ethnicity, or other protected characteristics. Hallucinations can perpetuate bias by generating stereotypes and fabricated facts or erroneous interpretations. Groups of people may fall victims of hallucinations and suffer severe consequences. Hallucinations can erode the protective environment of institutions and authorities that deploy them. If government bodies, insurance companies, healthcare authorities and financial institutions make critical decisions based on incorrect outputs the foundation of trust in institutions will be threatened. If hallucinations are left unchecked, they can contribute to the distortion of economic or democratic processes[19].

Various forms of biased patterns can display on LLMs as a result from the model's training set of data or design methodologies.

Data Bias occurs when the training data reflects societal inequalities. If a model is trained on data that over-represents one group or perspective, it will likely produce outputs that reflect those same biases like an echo-chamber where only identical opinions are expressed. Data bias may emerge as Social biases when models replicate historical and cultural stereotypes that reproduce inequalities based on gender, race, age, social status etc. Language biases arise when models favor certain languages, dialects, or linguistic forms, leading to uneven performance in translation and communication tasks. Such disparities risk deepening existing inequalities across education, commerce, and access to information by demonstrating preferential treatment or different languages and dialects[20].

Algorithmic Bias arises from the design and functioning of the algorithms that process the data. The way an LLM's architecture interprets and prioritizes information can unintentionally introduce bias. It doesn't matter how diverse the data set is. The model prioritizes certain patterns in data that reinforce stereotypes and discriminatory

---

[18] EU Commission High-Level Expert Group on Artificial Intelligence ETHICS GUIDELINES FOR TRUSTWORTHY AI, April 2019

[19] Socol de la Osa DU, Remolina N., Artificial intelligence at the bench: Legal and ethical challenges of informing—or misinforming—judicial decision-making through generative AI. Data & Policy. 2024; doi:10.1017/dap.2024.53

[20] Guo Y et al, Bias in Large Language Models: Origin, Evaluation, and Mitigation, arxiv, November 2024 available at https://arxiv.org/html/2411.10915v1

behaviors[21]. However not all biases have the same impact on natural persons. There are functional bias that are design choices that help improve model performance by adopting a structured reasoning process or consistent style. Some biases don't harm users but can limit the variety of responses. In contrast there are problematic biases that must be addressed and mitigated. They create outputs that unfairly disadvantage certain groups[22].

In order to overcome ethical challenges providers and deployers must reinforce constant scrutiny, ethical oversight, cultural sensitivity and value-based design and adaptation to technological and societal change.

Despite legal obligation for transparency regarding the interaction with AI, the possibility of erroneous outputs, the data used in the training of the system many LLMs do not warn users about the risk of hallucinations in a coherent way. Transparency is fundamental right of LLMs users. Lack of substantial explanation regarding the system and its possible limitations hinders their rights to be informed on the risks involved.

Since training datasets possibly contain personal data, there is an ethical duty of LLMs providers and deployers to take appropriate measures that ensure that this information will be safe and that appropriate measures will be taken in case a privacy breach occurs. Technical and organizational measures should be adopted to mitigate hallucinations: Augmenting and diversify training datasets to lower the rate of erroneous outputs[23], Reinforcement Learning with feedback that relies on human evaluators to review responses and finetune the system (enhanced RLHF).[24]

Hallucinations in LLMs represent more than just a technical flaw—they raise serious ethical concerns related to truthfulness, accountability, fairness, and public trust. Tackling this issue requires collective responsibility: developers and providers must adopt risk management strategies, professionals must ensure proper oversight, and policymakers must establish frameworks that safeguard users. Treating hallucinations as an ethical challenge rather than merely a programming bug is essential to ensuring AI contributes responsibly to society.

The policy debate around AI regulation centers on how to balance innovation with ethical safeguards. While oversight is necessary to prevent harms such as bias, misuse,

---

[21] Chen, Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanit Soc Sci Commun 10, 567 (2023). https://doi.org/10.1057/s41599-023-02079-x

[22] Bridewell W et al, The Technology of Outrage: Bias in Artificial Intelligence,arxiv, September 2024 DOI:10.48550/arXiv.2409.17336.

[23] Peng B et al, "Securing large language models: Addressing bias, Misinformation, and Prompt Attacks, arXiv, October 2024, arXiv:2409.08087 [

[24] Xiaoyu L et al, Integration of LLMs and the Physical World: Research and Application ACM-TURC '24: ACM Turing Award Celebration Conference 2024, Changsha, China, July 2024 DOI: https://doi.org/10.1145/3674399.3674402

and security threats, overly rigid regulation risks stifling technological progress and competitiveness. Experts recommend measures including explainability standards, ethics-by-design approaches, proactive compliance audits, dedicated governance structures, and continuous bias monitoring. Core regulatory principles emphasize transparency, fairness, and explainability, implemented through a risk-based model that aligns safeguards with potential harms. Policymakers are also weighing broader issues such as cybersecurity, malicious use, the establishment of oversight bodies, and international coordination to avoid regulatory fragmentation and encourage global cooperation[25].

Additionally hallucination detection mechanisms that identify hallucinations at the token level and correct inaccuracies during generation are important to recognize and rectify inaccurate content. LLM hallucinations are more than a technical limitation. They challenge fundamental ethical principles of truthfulness, accountability, justice, and trust. Addressing them requires shared responsibility: developers and deployers must implement risk management practices, professionals must exercise oversight, and policymakers must create frameworks that protect users. Only by confronting hallucinations as an ethical problem—not just a programming flaw—can ensure that AI systems serve the public good responsibly.

Conclusion

Hallucinations in Large Language Models (LLMs) are more than a technical defect: they expose users to risks of misinformation, bias, discrimination, and erosion of trust in institutions. Because LLMs are increasingly deployed in high-stakes domains such as law, healthcare, finance, and education, the consequences of erroneous outputs can be severe—affecting not only individuals but also democratic processes and social cohesion.

The comparative analysis of the European Union and the United States highlights two distinct philosophies of governance. The EU has adopted a rights-centered, precautionary framework through the AI Act and the GDPR, emphasizing risk classification, transparency, and accountability. Yet, hallucinations are not explicitly addressed, and responsibility remains diffuse, particularly in terms of ex post remedies for individuals harmed by erroneous outputs. The U.S., by contrast, relies on a fragmented, innovation-driven model, combining sector-specific laws, state initiatives such as Colorado's AI Act, professional standards, and voluntary guidelines. While this fosters flexibility and competitiveness, it leaves uneven protections and regulatory gaps.

Despite these differences, both jurisdictions converge on key regulatory priorities: transparency, explainability, risk-based oversight, and impact assessments. What is missing is a direct recognition of hallucinations as a regulatory concern. Treating them only as incidental flaws risks leaving systemic harms unaddressed.

---

[25] Olatokun W, Hallucination in the era of generative AI : Challenges to data integrity ethical concerns and implications to knowledge management, The 2025 DLIS 5 Biennial International Conference Strengthening Information and Knowledge Management Paradigms in the Era of Advanced Technologies, April 2025

Moving forward, effective governance requires a hybrid, multi-level approach. Technical interventions—such as diverse training data, reinforcement learning with human feedback, and token-level hallucination detection—must be combined with legal obligations for risk management, accountability, and user protection. Developers and deployers should share responsibility, while professionals must exercise due diligence in fields where errors have critical consequences. At the same time, international coordination between the EU, the U.S., and other jurisdictions is essential to prevent regulatory arbitrage and ensure consistent safeguards.

Ultimately, hallucinations must be seen not merely as a programming flaw but as a sustainability challenge for AI governance: how to balance innovation with the protection of human rights, fairness, and democratic values. By embedding ethical principles into both design and regulation, policymakers can ensure that LLMs strengthen, rather than undermine, public trust. Only then can artificial intelligence be harnessed responsibly for the public good.