

Ethical Risks and Safeguards in Generative AI: Managing Impacts on Cognition and Decision-Making

Risques éthiques et garde-fous de l'IA générative : maîtriser les impacts sur la cognition et la prise de décision

Houda BOUMEDIENE¹ 

AILE Laboratory, Amar Telidji University of Laghouat/ ALGERIA
h.boumediene@lagh-univ.dz

Ouafa OUARNIKI² 

SIGMA Laboratory, Ziane Achour University of Djelfa/ ALGERIA
ouafa.ouarniki@univ-djelfa.dz

Nadežda STOJKOVIĆ³ 

University of Niš/ SERBIA
nadezda.stojkovic@elfak.ni.ac.rs

Received: 10/04/2025,

Accepted: 07/10/2025,

Published: 25/12/2025

Abstract:

This study concentrates on the double-edged nature of AI as an innovation enabler and a source of ethical issues such as hallucinations, disinformation, and output variability and brittleness of large language models (LLMs) to prompts and context. As AI increasingly informs human decision-making and can amplify existing cognitive biases, it also provides the opportunity for improvement. The paper calls for a firm moral foundation for the development of AI to guide innovation toward societal values. Proper use of AI requires developers, researchers, policymakers, and end-users to work together. Ongoing end-user engagement and training are essential in managing risks as well as establishing the best practices. Legal frameworks will be needed to keep up with the faster growth of AI. Openness, ethics, and transparency must be given priority to usher in human development and ensure that AI drives innovation according to ethical principles.

Keywords: Artificial Intelligence AI-ethical frameworks- large language models LLMs- sharing of experiences- societal values

Résumé :

Cet article explore la double nature de l'intelligence artificielle (IA), qui peut être à la fois source de défis éthiques majeurs et catalyseur d'innovation. Parmi les principales préoccupations figurent les hallucinations causées par l'IA, la désinformation et la variabilité des sorties et la fragilité contextuelle des grands modèles de langage (LLM). Bien que l'IA informe de plus en plus la prise de décision humaine et puisse amplifier les biais existants, elle représente également un potentiel de progrès considérable. L'étude souligne la nécessité d'un cadre éthique solide pour guider le développement de l'IA tout en conciliant innovation et valeurs sociales. Pour une utilisation responsable, la coopération interdisciplinaire entre chercheurs, développeurs, décideurs politiques et utilisateurs est cruciale. Pour réduire les risques et diffuser les bonnes pratiques, le partage d'expériences utilisateurs, l'implication active et la formation continue sont essentiels. De plus, les cadres juridiques et réglementaires doivent évoluer au rythme des avancées technologiques. En favorisant transparence, éthique et dialogue, l'IA peut contribuer positivement au progrès humain tout en respectant les valeurs d'éthiques.

Mots- clés : Intelligence artificielle (IA) – cadres éthiques – grands modèles de langage (LLMs) – partage d'expériences – valeurs sociétales.

* **Corresponding author: Houda BOUMEDIENE**

Langues & Cultures / © 2025 The Authors. Published by the University of Adrar, Algeria.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

Introduction

Artificial intelligence (AI) has become an innovative technology that has quickly transformed healthcare, education, finance, transportation, and other industries. The progress race, especially with large language models (LLMs), has opened doors to innovation at a scale and magnitude never experienced in previous waves of technology (Bommasani et al., 2022). These technologies come with daunting challenges, particularly the dissemination of disinformation, the spread of bias, and the phenomenon of AI hallucinations (Xie & Wang, 2023).

AI hallucinations, assured but deceptive responses, can result in catastrophic mistakes: computational errors in financial processes, software bugs in autonomous systems, or false theoretical connections in clinical reasoning (Burtell & Woodside, 2023). Their risk is amplified by the fact that they often appear amidst otherwise accurate data and are not always easily detectable or correctable by human users. Placing these dangers in perspective requires tracing the rapid development of AI from rule-based systems through deep learning to modern neural architecture. As capabilities become more sophisticated and autonomous, the potential and magnitude of unforeseen consequences increase as well (Zhang et al., 2023). Documented issues, ranging from biased medical recommendations to racist facial recognition, underscore the urgent need for conscientious design, deployment, and oversight (Buolamwini & Gebru, 2018; Raji et al., 2020).

Another concern is the nature of human-AI interaction with cognition. System designs may either overlook cognitive vulnerabilities (e.g., anchoring and confirmation effects) or reinforce them, thereby producing technical discrimination without adequate socio-ethical protections (Tversky & Kahneman, 1974; Kahneman, 2011). Socioeconomic damage also persists among marginalized groups, such as when biased identification systems produce disproportionate legal sanctions or artificial financial activities distort markets and erode trust (Alkaissi et al., 2023).

Socioeconomic damage also persists with marginalized groups, such as when biased identification systems produce disproportionate legal sanctions or artificial financial activities that cheat markets and destroy trust (Alkaissi et al., 2023). Precautions should ensure maximum quality, user control, and the integration of complex perspectives from AI system cradle to grave (Gebru et al., 2021; Mitchell et al., 2019). Regulatory and juridical frameworks have to keep up with rapid-paced innovation in an attempt to provide social and moral standards (Klenk, 2024). Ongoing learning, engagement, and appropriately structured sharing of users' practice can expose recurring issues and validate best practices for ethical uptake (Xie & Wang, 2023). Because of the accelerated rate of change, alignment with current AI research and engineering advancement becomes priority to enhance reliability and safety. Progress relies on researchers, developers, policymakers, and end-users collaborating in concert. Open

communication, sharing of knowledge, and clear ethical commitments are required to achieve benefits with minimal harm (Floridi & Cowls, 2019; Klenk, 2024). With critical thinking and ethics-conforming deployment, the transformative power of AI can be used for significant human and societal benefits.

1. Research Design

We looked at peer-reviewed literature and standards or policy papers published between 2015 and 2025 on large language model (LLM) failure modes such as hallucination and output brittleness, human-automation trust and decision-making, persuasive and choice-architecture features in digital technologies, and broader discourse on AI ethics and regulation. Backward and forward citation chasing supplemented the searches to capture all relevant angles. For each source, we noted the AI affordance or breakdown mode, the implicated cognitive mechanism, the ethical principle addressed, the use context (e.g., education or healthcare), and any controls or safeguards that were mentioned. Codes were iteratively refined for conceptual consistency and coherence across the corpus. The derived data were then combined into a mapping framework connecting AI affordances to cognitive mechanisms, aligned ethical risks, and suggested controls, which we categorized under design, process, governance, and user practice. Two illustrative vignettes, one on healthcare triage and another on student writing support—are employed to show how risk-control pathways function in practice.

In our attempt to introduce rigor and prevent bias, we double-checked coding on a pilot set, drew evidence from a wide range of fields including computer science, HCI, psychology, and ethics/policy, and made claims only where convergent evidence was available, thereby avoiding over-generalization. Ethical analysis was guided by the principles of autonomy, non-maleficence, justice, and accountability. No human subjects were involved. The main limitations of this review concern the rapid pace of AI development and the restricted access to proprietary platforms, which may affect completeness, as well as heterogeneity in the evidence base, which constrains causal inference. Accordingly, the safeguards presented here should be considered theory-informed propositions rather than definitive conclusions.

2. AI Hallucinations in Critical Sectors: Challenges and Solutions

One of the major challenges in the deployment of AI systems, especially language models like ChatGPT, is a tendency to generate responses that might sound plausible but are factually wrong or even fabricated. Such deviations make a huge difference in areas that demand great accuracy and reliability, such as scientific writing and medical education. They are a product of the training methods of LLMs, including BERT, ChatGPT, Claude, LaMDA, and Llama, trained on large datasets containing all types of information, including incorrect or misleading data; hence, this inaccuracy sometimes reflects in their output. In contrast, Schemmer et al. (2022) explain one of LLMs' most substantial weaknesses: unmarked error generation, from mathematical and coding errors up to misattribution and conceptual disambiguation. These hallucinations, which are popularly called so, may be impossible to

separate from otherwise correct information; therefore, they are difficult to identify unless in vigorous examination and diligent fact-checking.

In closed-domain settings, LLMs have tried consistency tests and may present inconsistencies beyond facts and information provided. This occurs when AI models generate out-of-context responses or drift away from context data to provide incorrect or meaningless outputs. Although closed-domain hallucinations are limited, e.g., overfitting training data, bias, or unrelated responses, they are relatively more detectable and correctable since the knowledge domain covered is finite. Open-domain hallucinations, however, are worse as they generate spurious or misleading content on any subject without any context that has been provided. They involve the generation of factually incorrect information, illogical text, biased text, or totally unrelated text, all because AI is based on patterns learned and does not comprehend the subject. Handling such hallucinations in open-domain AI systems that must explore very extensive and varied knowledge landscapes remains extremely challenging.

The work of Insiew & Alfandi (2024) delves into the complex interplay between human trust and AI performance, topics particularly relevant within sectors such as health, law enforcement, and autonomous systems, where AI performance is a critical consideration. It underlines how inconsistent or flawed AI performance undermines human trust, a crucial factor in high-stakes domains wherein AI recommendations or decisions strongly influence operations. It also points to complex dynamics in human-AI interaction, where perceptions of AI (as human-like or distinct) make a great difference in trust and collaboration. These areas require very high accuracy and reliability, whether for medical diagnostics, decision-making in the legal sphere, or autonomous navigation, and the consequences of trust loss due to AI malfunction or "hallucinations" are serious. The present body of work therefore demands AI systems that provide consistent performance, allowing users to build trust and avoid risks due to AI failures, so that AI augments rather than hinders important decision-making.

2.1. AI Hallucinations in Data and Content

AI hallucination is not only confined to generative content that generates material or text but also to processing and interpretation through an AI system. Most apps need AI to sort through an enormous ocean of information to come up with an insight or make a decision. However, "if AI hallucinates, or generates misleading or entirely invented output, it severely tarnishes the reputation of such an insight or decision" (Eysenbach, 2023). For instance, in medical or financial applications where accurate interpretation of data is paramount, hallucinations by AI can lead to erroneous diagnoses, erroneous financial projections, or other pivotal errors. This analysis can cause risks to people or businesses by virtue of AI-driven decisions.

To avoid such risks, AI systems must be trained on high-quality data and must have the ability to detect and correct hallucinations. This can be through robust validation methods that attempt to test AI output for validity. Incorporating domain-expert feedback in training AI models can also render interpretation of such intricate datasets more efficient (Wang et al., 2020).

Transparency is another key consideration. AI decision-making needs to be transparent: the process through which AI arrives at a decision and on which sources of data it has based its decision needs to be traceable to create trust in AI-powered systems. Users, whether health

experts, financial analysts, or end-users, need to understand how AI arrives at conclusions so that they can work with its results safely.

The impact of AI hallucination is far-reaching and can be observed in the equipment used by the world, such as content moderation on YouTube. When AI does not interpret or type information, it may end up leaving out risky or obscene material. Thus, vigilant monitoring and regulation of AI machines are required. Inevitably and continually updating ways to distinguish correct information from hallucinations can help avoid the risks involved in AI-driven decision-making processes in all sectors.

2.2. Key Steps towards Protecting the Reliability and Credibility of AI Systems

The AI hallucination process is an intrinsic issue that deserves genuine and serious consideration. Honesty and accuracy of AI-generated content are especially of extreme importance in genuine contexts such as medicine and academia. Although there are growing studies being conducted on making AI systems more robust and needing transparency, the AI community can surmount the hallucination issue and widen application reliability. Longitudinal monitoring by Carroll et al. (2023) underscores the intricacy of the issue and the need to apply various methods of solving it in an overarching manner.

Some effects emerge when AI inaccurately transcribes insensitivity content, a scenario seen on platforms such as YouTube. Such illustrations convey interpretive errors and the possibility of AI spreading misinformation or objectionable content with a domino effect on public opinion (Zhang et al., 2023). This is a reason to introduce more stringent content moderation and reality checks.

In fact, addressing hallucination needs a sustained focus on robustness and transparency. Through collective insights and effective implementation, the AI community can reduce risks, enhance reliability, and promote responsible use across varied applications.

2.3. Hallucination Risks in Crucial Contexts

As AI is increasingly used in health care and other domains, its potential for sophisticated work, and sometimes more than human capabilities, has been witnessed. Nevertheless, AI hallucinations and deceptive or misleading results contradict radical ideas about trust and dependability, most of all for high-stakes applications like health care. In high-risk environments, AI errors or recommendations in suggestions or conclusions can lead to severe, even fatal, impacts. Confronting this challenge requires AI that is accurate, transparent, and accountable, along several lines:

- **Accuracy:** High-quality and diverse datasets are needed for training, so the model avoids egregious errors. Upon strict verification, regular monitoring should ensure the accuracy of AI-produced content.
- **Transparency:** Systems must be transparent, stating how conclusions were reached and with sources referenced for users to authenticate.
- **Accountability:** Users must be able to critically assess AI output, challenge mistakes or prejudice, and acquire recourse. A culture of transparency and

accountability helps users to recognize and discard hallucinated content more effectively.

These strategies are crucial. In medicine, for example, AI capabilities in diagnosis, prescription, and management can revolutionize outcomes—but the situation could worsen if systems churn out incorrect information. Maintaining reliability and transparency maintains integrity of outputs and safeguards patients and users. With systems that are more interpretable and reliable, goodness of AI is within our grasp while risks of hallucination are diminished, an equilibrium for high-stakes domains where high accuracy and reliability are needed.

2.4. Strategies for Enhancing LLM Reliability

The sudden integration of AI in a high-risk arena such as finance and healthcare reflects the capability of AI to accomplish complex tasks, often even better than humans. This triumph, however, has its price in the form of the problem of AI hallucinations, where systems generate misleading or made-up results. In high-risk environments, the implications are catastrophic. Overcoming such a challenge requires a comprehensive effort to improve LLM robustness and reduce hallucination risk. Carroll et al. (2023) suggest a way to manage risk through expand training data with accurate and reliable information. Ensuring data is comprehensive, complete, and factually accurate reduces the hallucination incidents. No less important are rigorous validation checks during development and deployment. Active validation and monitoring against known facts and domain knowledge enable systems to detect and correct hallucinations prior to their presentation to end-users. Expert feedback is also important : domain experts who provide commentary on data subtleties and facilitate the capture of errors that cannot be captured through automation. Additional work with experts permits ongoing improvement and more accurate interpretation and creation.

Transparency underpins trust and accountability. Models can be trained to record lines of reasoning and provide checkable sources so users can separate supported facts from potential hallucinations. At the same time, accountability culture within the AI community can enforce best practices. In the face of the risk of misinformation—especially on large platforms—more and better content-moderation tools are increasingly necessary. Sophisticated tools can ensure compliance with factual information and ethical norms. Finally, development and vigilance are reliability buzzwords. Developing models, through continuous updating, performance testing, and incorporation of novel findings, can minimize hallucination risks. (Ferreira & De Souza, 2021). Users of LLMs are also served by ongoing education regarding model constraints and by competencies to distinguish fact from hallucination. This kind of literacy, along with transparency and accountability, functions to guarantee AI output integrity.

Furthermore, accuracy of inference is perhaps less critical for creative or exploratory LLM uses (e.g., fictional writing support). In those cases, hallucinations may be acceptable if there is explicit grounding material and an intensive end-user review cycle (Ferreira & De Souza, 2021). In contrast, in high-stakes domains, where truth and accuracy are non-negotiable—strong vetting of AI output is imperative. Unrecognized hallucinations can propagate downstream and contaminate future datasets and training.

Extreme caution and extensive review are therefore advised in fields such as medicine, journalism, transportation, and any situation where language or behavior is attributed to individuals or

organizations. Early uses of ChatGPT in technology journalism, for example, led to serious mistakes in published work; new review processes were subsequently instituted when LLMs entered writing workflows.

2.5. Strategies toward Robust and Ethical Applications

Although AI is of enormous potential in the majority of fields, tackling hallucinations is central to ongoing safety, reliability, and credibility. It is imperative that the AI community assign an utmost value to training-data quality, improve error-detection mechanisms, and impose greater transparency in order to curb hallucinations and safeguard the integrity of AI-generated content. There will be a requirement for all concerned—researchers, developers, and consumers, to exchange challenges, best practices, and standards impacting AI in major domains. Moral principles and regulatory mechanisms need to evolve so that appropriate boundaries can be set such that AI systems not just remain technically right but are also morally aligned with social values and norms. For those sensitive areas, regulatory bodies may have to be more interventionist so proper testing and certification can be done for high-risk use.

Education and awareness are equally important. Stakeholders should be informed about AI limitations and potential risks. Health professionals, scientists, and content producers, for example, can all gain by learning to identify AI-generated content and how to apply critical, evidence-based analysis. The long-term vision entails that AI systems are embraced as reliable supplements to human judgment and help to augment, not degrade, decision-making (Wang et al., 2023; Yu et al., 2018). Continued study of the causes of hallucinations, their remedies, and mitigations will permit AI to realize its potential for good while protecting it from harm. It is a way that demands both vigilance and active cooperation on the part of everyone involved.

3. Protecting Decision-Making from AI Manipulation

3.1. Autonomy and Personal Agency

The increasing role of AI in human decision-making has emerged as a high-priority issue at the intersection of technology, ethics, and human nature. Beneath all of these issues lies diminished human agency with AI systems, and especially those fueled by predictive recommendations and decisions, taking a central function in human decision-making. The blending of AI capability and human autonomy calls into question the ethics of free will and individual agency.

3.2. Everyday Integration and Societal Shifts

As AI technologies shift and change, they are increasingly being incorporated into day-to-day decision-making—ranging from customized content streams to multi-dimensional business strategy. This will mean a rethink on machine input, giving way to outcomes and potentially even replacing personal judgment. The effect is insidious but pervasive: aside from people, mass behaviors and social norms could well shift under algorithmic inputs.

3.3. Tactics and Moderators of Influence

Recent research has investigated mechanisms by which AI systems influence human decision-making and preferences. These involve personalization using user data—to make individualized recommendations—and confabulation, in which models generate realistic but false information that may unintentionally mislead users. Recommender systems can also direct preference towards more extreme content to generate high engagement. Indeed, the subtlety of AI's influence, the perceived intentions behind a system, and potential harms all moderate its effectiveness in shaping decisions. People may be unaware that an external agent is shaping their beliefs; moreover, users judge systems differently depending on whether they perceive the AI's intentions as aligned and benevolent (Schemmer et al., 2022).

3.4. Detection, User Models, and Mitigation

Everitt et al. (2021) state that AI manipulation detection relies on the interaction of system design and transparency, users' knowledge and mental models of AI, and user traits like trust in technology. Cognitive capabilities and AI experience also influence the detection of manipulation. The mitigation approaches include greater transparency and explainability, adaptive interaction for sensitivity to user knowledge and preferences, ethics-by-design, feedback systems, cultural and contextual sensitivity, safety and reliability considerations, and shared decision-making. Through its impact on human decision-making, AI has long-term implications for fairness and bias, scientific progress and conflict, power and inequality, epistemic processes, human rights, and ethics.

3.5. Balance and Governance Imperatives

Human agency and AI are not a subject of theory; the effects are real. There must be a balance between unleashing the potential of AI without undermining human volition. As we cross this divide, there must be ethical foundations and models of governance put in place so that AI enriches and does not debase the human experience. This requires an inclusive response—a combination of technical solutions, ethical norms, regulatory frameworks, and stakeholder involvement—founded on transparency, user empowerment, and aligning AI development with human values.

3.6. Forward View and Operationalization

As AI comes to shape decision-making more generally, ethical and societal hazards call for reflective and proactive intervention. By foreseeing manipulation tactics, moderator variables, and downstream consequences, we can architect AI technologies to promote human flourishing and drive an equitable, sustainable future. To operationalize these orientations in practice, Table 1 charts AI affordances onto the psychological processes they activate, the ensuing decision vulnerabilities, and ethically warranted safeguards and metrics.

Table 1. Ethical–Psychological Pathways of AI Influence.

AI Affordance/ Pattern	Psychological mechanism	Decision risk/ outcome	Ethical principle implicated	Safeguard/ control	Suggested metric
---------------------------	----------------------------	---------------------------	------------------------------------	-----------------------	---------------------

Ethical Risks and Safeguards in Generative AI: Managing Impacts on Cognition and Decision-Making.

Personalization & ranking	Anchoring; peripheral-route acceptance	Choice shifts; over-reliance on defaults	Autonomy; Justice	Exposure diversity controls; brief explanations; opt-out	Exposure diversity index; override rate of default
Plausible-but-false outputs (hallucinations)	System-1 acceptance; automation bias	Error propagation into actions/workflows	Non-maleficence; Accountability	Uncertainty cues; citations; second-source check; human-in-the-loop	Error rate vs. ground truth; citation coverage
Missing provenance (retrieval/attribution gaps)	Confirmation bias; reduced scrutiny	Low verifiability; misleading summaries	Accountability; Autonomy	Content provenance (e.g., C2PA); source links; compare-view	Share of outputs with verifiable sources
Affective targeting (emotion-aware messaging)	Arousal-driven reduced scrutiny	Emotion-driven decisions; regret	Autonomy; Non-maleficence	Label persuasion; throttle intensity; “plain mode” option	Accuracy vs. arousal trade-off; user complaint rate
Algorithmic bias in models/data	Stereotype reinforcement; availability	Disparate impacts across groups	Justice	Bias audits; subgroup thresholding; fairness monitoring	Subgroup performance deltas; disparity indices
Nudging via defaults/salience	Default effect; salience bias	Path dependence; reduced deliberation	Autonomy	Resettable defaults; confirm-to-proceed; side-by-side alternatives	Default uptake variance after reset/friction

4. Understanding AI Manipulation Techniques

The AI systems use a set of subtle techniques to influence human behavior and decision-making. These techniques are useful but equally present several ethical and practical problems. The section below depicts those specific techniques that, as per Bubeck et al., (2023), the AI systems use in influencing human choice and provides examples and case studies relevant for explaining such techniques in real life.

4.1 Techniques of Personalization and Recommendation Systems:

Data Collection and Analysis: AI systems gather a huge amount of data from the users regarding their browsing habits, social media interactions, purchase history, among others. This data is further analyzed with a lot of detail in order to develop really elaborative user profiles. Based on the consumer profiles created, AI algorithms personalize news recommendations, videos, products, and services for users depending on individual preferences.

Examples and Case Studies:

- Netflix and Spotify also use AI in recommending movies, TV serials, and songs that the particular subscriber would like. Recommendation algorithms take note of the users' past viewing or listening history, ratings, and even times of day they are most active.
- Amazon: An e-commerce giant such as this one employs AI so as to make certain product recommendations, taking into account the history of a user's browsing and purchasing. Indeed, while this type of personalization enhances user experience and increases sales, it does raise concerns regarding consumer manipulation and its potential to encourage overconsumption.

4.2. Techniques for Producing Plausible but False Information

- **Language Models:** Advanced models like GPT-4 produce more coherent and contextually accurate text, but are more prone to generating information that sounds incredibly plausible but is partially or wholly made up
- **Contextual Relevance:** AI-generated content is prone to simply fitting in with the context of the question or conversation and thus becoming hard for the user to recognize errors.

Examples and Case Studies:

- **AI Chatbots:** AI chatbots in customer service create responses which seem to be truthful but can be misleading. Simple examples could be a chatbot supplying false information concerning the availability of a product or giving the wrong measurement of an item.
- **Deepfake Technology:** This is an AI-generated video that makes it quick and realistic to create false images of people that could manipulate public opinion and lead to the misinforming of society.

4.3. Behavioral Targeting and Nudging Techniques:

- **Behavioral Analysis:** AI programs examine the behavior of users and search for patterns so that they can anticipate what the user is likely to do next.
- **Nudging:** Using AI, a system could nudge the users' decisions slightly by rephrasing the choices so that one would obviously appear to be better than the rest. To achieve this, it could reorder search results, emphasize certain choices, or do other manipulations of the same kind.

Examples and Case Studies:

- **Google Search:** The results are ordered by the search engine through AI, which identifies which results are more relevant to a given query and based on users' browsing behavior. By ranking some results higher than others, Google influences what results users are most likely to see and believe to be true.
- **Social media:** On platforms like Facebook and Twitter, algorithms powered by AI sort feeds and recommend similar content. This reinforces the very same ideas that were the object of search queries and builds an echo chamber in which thought patterns might come to shape public opinion and behaviors.

4.4. Affective Response Techniques Manipulation:

- **Sentiment Analysis:** AI systems analyze text and speech to check user emotions and adapt to them.
- **Emotional Engagement:** AI, through emotionally charged language and imagery, elicits certain emotional responses from users and influences them in making certain decisions and taking certain specific actions.

Examples and Case Studies

- **Advertising Campaigns:** AI-powered advertising platforms use sentiment analysis to concoct emotionally engaging advertisements; these Advertisements will build emotional resonance with users' feelings and desires, which may be used to provide a high likelihood of making a purchase.
- **Political Campaigns:** AI systems in political campaigns compose messages that could have an appeal to citizens emotionally and manipulate their opinions and voting behavior.

4.5. Algorithmic Bias and Fairness Techniques:

- **Data Bias:** The AI system propagates biases within the data upon which it has been trained. The result of this process will be biased, or unfair.
- **Algorithmic Decision-Making:** Decisions by AI algorithms are informed by past bias and reinforce it. Making decisions based on facts regarding the past may impact employment, lending, or law enforcement.

Examples and Case Studies:

- **Predictive Policing:** Police technologies that employ artificial intelligence have been repeatedly criticized, among other things, for disproportionately targeting minorities due to their biased training data.

- **Hire Algorithms:** Various firms utilize AI to sift through job applicants. These algorithms have been seen as discriminating against certain classes of people in favor of others and raise concerns about how fair and transparent the AI systems will be in implementation.

4.6. Influencing Public Opinion and Social Norms Techniques

- **Amplification of certain content types:** AI can amplify some types of content to greater visibility and influence in public discourse.
- **Creation of Echo Chambers:** Because of AI-driven personalization of content, users exist in a space where exposure to information could only reiterate views that they already hold.

Examples and Case Studies:

- **Twitter Trends:** AI-powered trending on Twitter may shape public discourse by bringing certain issues to the front stage from backstage.
- **YouTube Recommendations:** The recommendation algorithm of the site has been known to push users toward progressively more extreme content, shaping public opinion and perhaps the shifting of social norms

5. Challenges of Managing the Impact of Deceptive Generative AI on Human Perception and Decision-Making

5.1. Technical Difficulty and Detection

- **Simple Generative Models:** Generative AI models have become more sophisticated; models like GPT-4 develop output that is often very convincing yet, in fact, deceptive. The difficulty is not just in hallucination and fabrication but that the information may be juxtaposed against information that is true. The problem, therefore, is the complexity in trying to devise systems that can detect and correct this miscommunication in real time.
- **Limited Explainability:** Most of the generative models are black boxes that provide output without explaining how the output has been derived. Since there is a lack of transparency in internal mechanisms, it is hard to comprehend and determine when and why these models generate misleading information, and it is hard to control and mitigate their consequences.

5.2. Quality and Bias in Data

- **Problems with Training Data:** Generative AI models are typically trained on large volumes of data, which may be biased, inauthentic, or misstated. These models inadvertently pick up those biases and thus can produce biased outputs that might also be harmful in perpetuating stereotypes or false narratives.
- **Diverse and High-Quality Data:** The major challenge is that an AI system should be trained using diverse and high-quality datasets. The scale of data required for these models is so high that, after careful curation, almost a certain degree of bias or misinformation is likely.

5.3. Ethical and Moral Dilemmas

- **Manipulation and Autonomy:** There is a chance that generative AI could subtly influence human perception and behaviors, raising ethical questions regarding the role of manipulation versus respect for autonomy. The trick will be in developing systems that honor user self-determination while delivering valuable insight without overstepping their ethical bounds.
- **Ethics and Innovation:** While the utilization of AI capability has a thin line in dividing the benevolent use from preventing misuse that might lead to deceiving, there is, therefore, a need for managing the balance through the ethical framework in AI development and deployment without stifling innovation.

5.4. Regulatory and Legal Hurdles

- **Lack of Comprehensive Regulation:** Since the AI generative technologies keep evolving on a day-to-day basis, the regulatory mechanisms have been lagging behind the innovations. It would be hard to come up with regulations and rules that would prove to be useful and helpful in managing the use of AI, most notably in its generation and spread of misleading information.
- **Global Coordination:** The drafting of regulations that apply on a global scale is another huge challenge. Different countries have their own perspective on privacy, free speech, and how AI technology is used ethically; thus, making something that applies universally is tough.

5.5. Trust and Transparency

- **User Trust Loss:** The authority granted to AI, which produces counterfeit content, destroys users' trust in the systems. Trust defeats the speed of deployment of AI technologies and minimizes the effectiveness of technology use in decision-making. Regaining this lost trust is important to ensure responsible AI.
- **Challenges to Transparency:** In the process of the operation of AI systems, it will be quite hard to introduce transparency especially into the generative models whose output can be generated from big and variant input data. While users cannot understand how AI reaches an output, they cannot trust them enough for effective use. These remain challenging to perform in such a way that they would not overwhelm users.

6. Mitigating Unintended Consequences

- **Unforeseen Outcomes:** Implementation of AI systems in practice has, whenever possible, meant a number of unintended consequences, such as amplifying false narratives and reinforcing biases. These are hard to anticipate, let alone mitigate; the difficulty keeps increasing with increased AI autonomy and growing diversity of use cases.
- **Dynamic Adaptation:** Since generative AI is a domain that keeps constantly changing, it means mitigations or strategies against this technology have to be dynamic, plugged in for

monitoring at all times, evaluation, and readjustment of AI systems so as not to allow the passing of deceptive content.

- **Lack of Knowledge:** Many people do not understand the capabilities and limits of generative AI, so they are easier to fool. Again, this turns out to be a basic challenge, namely the education of users on how AI may also generate misleading information.
- **Critical Thinking Promotion:** It has to do with achieving critical thinking and skepticism during the interaction with AI-generated content. However, it is hard to achieve on a large scale, since it deals with changes in user behaviors and perceptions about the reliability of AI.
- **Cost and Resource Constraints:** The comprehensive monitoring and mitigation strategies require substantial resources to implement. It thus constitutes the challenge for an organization with a squeezed budget. The balance between the requirements of adequate safeguards and the harsh realities of cost and resource management will persist.

Addressing these challenges means stakeholders must work toward a more ethical and responsible deployment of generative AI-that the benefits are maximized while potential harms are kept to a minimum. It would involve some degree of collaboration among developers, policymakers, researchers, and end-users in finding their way out of this complex landscape of ethics and regulation related to AI.

7.Discussion

7.1. Dual Nature of AI and the Need for Balance

This deep-diving paper in the vast landscapes of AI underlines some of the most key challenges: AI hallucinations, misinformation, and emergent, unpredictable behavior of conversational AI/search engines. The multi-pronged strategy is highly emphasized in this study, where collaboration among researchers, developers, policy makers, and users is needed to handle the above-mentioned challenges. The important understanding here is that all these AI technologies are inherently Janus-faced: inasmuch as they represent unparalleled opportunities for innovation and solution-finding, they are also equally important sources of risk, such as erosion of privacy, manipulation of human behavior, and dissemination of false information. This duality underlines the need to find a balance between exploiting the benefits derived from AI while mitigating its potential negative effects. All these reviewed studies do not give the required framework for this balance; hence, the need for further investigation to develop comprehensive guidelines and best practices in ethical AI development and deployment.

7.2. LLM Weaknesses and the Open Trust Problem

The weaknesses in LLMs that are very critical include the generation of errors without warning. This remains an enormous challenge in the hands of users to check whether what information is accurate or misleading. These studies note the need to introduce sophisticated techniques that would immediately identify and correct AI hallucinations. In no way, however, have they gone so far as to offer concrete proposals or comparisons of effectiveness in approaches. Future research is

needed to develop and test new methods for identifying and reducing AI hallucinations. A mix of several may be useful. The second essential area that has been dealt with in the current reviewed literature is the correlation between human trust and AI performance. While such studies investigate the decline in trust due to inconsistent or incorrect AI output, they fail to examine further variables leading to it and the long-term effects on the adoption and use of AI. Further research should outline the factors affecting trust of users in AI systems and elaborate strategies that will allow gaining this trust and maintaining it over time.

7.3. Ethical Implications, Psychological Vulnerabilities, and Malicious Manipulation

It can be used to affect human decisions and dominate psychological weaknesses; the ethical problems surrounding it are extremely profound, evoking core issues with privacy and mental status. However, none of these studies has looked deeply enough into these ethical implications and/or put forward any detailed guidelines with a view to addressing them. Further work may deal with evolving ethics and regulatory mechanisms which can govern the use of AI in sensitive areas and also take into account the repercussions of such regulations (Athaluri et al., 2023). Case studies on malicious manipulation of chatbots like Microsoft's Tay and Sydney and Eliza show how AI systems can be targeted for malicious ends, flagging off serious ethical, regulatory, and technical challenges to the deployment of AI in sensitive contexts. These examples send a signal that proactive measures are needed in developing AI; however, they do not study the efficacy of various strategies regarding countermeasures or the implementation challenges.

7.4. Explainable AI (XAI): Promise and Open Questions

Explainable AI therefore points very promisingly toward improving trust and transparency in AI systems. On the other hand, the mixed influence of XAI on decision-making performance, user trust, and perception suggests that much more research will be required to optimize the interaction between AI and human users. Such further investigations may follow questions such as: What factors determine the effectiveness of different XAI techniques? Are there potential trade-offs between explainability and other desirable features of AI systems, such as accuracy and efficiency?

7.5. Unpredictability: Detection, Metrics, and Standards

While promising, the detection, measurement, and mitigation strategies in the reviewed AI research on unpredictability still lack wider coverage with frameworks and guidelines that would prescribe their application in a consistent, yet practical way across a very wide variety of AI systems and situations. Further work in this area may focus on developing standards of metrics and benchmarks by which the actual performance of these strategies can be evaluated, as well as the synergies that can be realized when many approaches are combined (Eysenbach, 2023).

7.6. Limits of Current Evidence and Directions for Robust Study

The limitations of the reviewed studies include a focus on specific AI challenges and domains that limit the generalisability of findings in other settings. Future work should, therefore, consider an

integrated approach to allow investigations into how such interlinkages occur across different AI challenges and their combined impacts on society. Further, many of the studies fall back on theoretical models and assumptions that arguably cannot capture the full complexity and dynamic nature of real-world AI systems. Further research might adopt a range of methods, such as empirical studies, simulations, and case studies that could give a more nuanced look at AI challenges and possible solutions.

7.7. Synthesis and Forward-Looking Agenda

The paper represents a valuable synthesis of the ongoing state of art in the realm of challenges emanating from AI, pointing out the indispensable necessity of a balanced and proactive approach in both the development and deployment of AI. At the same time, it identifies considerable deficiencies in the available literature, especially regarding a search for concrete solutions, actual efficiency assessments of various approaches, and ethical consequences of these AI technologies. With AI continuously evolving, such lacunas in collaboration need to be addressed by researchers, developers, policymakers, and end users in general for the development of systems in AI that would benefit humanity with minimal harm. This would enable us to exploit its monumental potential while its development and deployment are informed by principles of responsibility, transparency, and ethical integrity.

8. Conclusion

This research demonstrated the complex space of AI, focusing on the main challenges presented by the hallucinations of AI, misinformation, and randomness. It also discussed the transformational power of AI through examples like the GPT-4 model, which would amount to reinventing economies, industries, and even world politics. It has also exposed the profound risks of AI, i.e., its potential to erode trust, influence human decision-making, and exploit psychological faults unfairly. Why this matters is clear: these effects occur in high-stakes domains and, if unchecked, can erode autonomy, fairness, and accountability.

How it functions was highlighted through links between AI affordances (e.g., variable outputs, personalization, persuasive interfaces) and cognitive mechanisms (e.g., heuristics, anchoring, confirmation, automation bias), which together help explain why misleading content is accepted and acted upon. A review of the research points out that proactive steps in AI development include extensive testing, strong safeguards, and ethical bases, in addition to openness. These are most important in sensitive areas of applications related to mental health, where AI systems are designed with great involvement from domain experts and therefore must keep ongoing checks on safety and users' well-being. It only underlines how much is expected on ethical embedding and cognitive synergy in a world where AI will find an increasing presence in many features of daily life. XAI has emerged as a promising direction of research that may help build a trusting relationship and make AI transparent about its decision-making process. The literature also points to education and awareness related to the end users for their responsible interaction with AI. This would involve digital literacy training, educational courses regarding the gaps in AI, and inculcating a sense of criticism towards AI-generated content.

This paper recommends to adopt an integrated methodology that combines (i) robust policy frameworks embedding ethics into development and deployment, (ii) periodically updated best practices in quality assurance to anticipate and mitigate risks of hallucinations and misinformation, and (iii) user-facing safeguards (uncertainty cues, provenance, friction for high-impact actions) that align with ethical principles of autonomy, non-maleficence, justice, and accountability. As we continue to commit more and more resources and intellectual capital to developing AI, it's important that efforts toward technological breakthroughs be securely grounded in ethical principles and consider the values of society. This will require continuous coordination among users, developers, policymakers, and researchers in dealing with the challenges presented by AI as well as its potential for public good.

While phenomenally promising, AI also poses significant challenges to be met with imagination and cooperation. By cultivating a responsible, transparent, and ethically informed approach to the development and deployment of AI, we can create a future in which AI systems are trustworthy, reliable, and responsive to the best interests of humanity. The present paper reviews current research on challenges and opportunities from AI and underlines a number of areas in which further examination will be required. It is urgent to develop more effective methods to spot and correct AI hallucinations in real time and find their root cause.

It is also extremely relevant to investigate the efficiency of different remedies for overcoming the biases in AI and ensuring the development and deployment of AI systems in a fair and inclusive way. Investigations into the complex interplays of AI, human cognition, and societal biases – and the elaboration of extensive frameworks concerning the treatment of vulnerability in humans vis-à-vis manipulation by AI systems – become central for the protection of individuals and societies. Finally, investigation into new AI training paradigms, which naturally embed safety and ethical considerations, and development of metrics to evaluate AI systems' robustness against adversarial interventions are vital to building AI technologies that can be trusted to be reliable, trustworthy, and aligned with human values. In brief, AI is important because it has social stakes; it operates along understandable cognitive and ethical pathways; and it needs to be managed and directed by material checks that render its principles functional. By investigating these knowledge gaps and remaining open to an ongoing examination of the challenges and opportunities posed by AI, we can work toward a future in which that immense potential is harnessed in responsible and ethical ways, serving all of humanity.

Bibliography

- Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(3), e35179. <https://doi.org/10.7759/cureus.35179>
- Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15(4), e37432. <https://doi.org/10.7759/cureus.37432>

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., & Liang, P. (2022). On the opportunities and risks of foundation models. *arXiv*. <https://arxiv.org/abs/2108.07258>
- Bubeck, S., et al. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. *arXiv*. <https://arxiv.org/abs/2303.12712>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research* (Vol. 81, pp. 1–15). <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Burtell, M., & Woodside, T. (2023). *Artificial influence: An analysis of AI-driven persuasion*. *arXiv*. <https://arxiv.org/abs/2303.08721>
- Carroll, M., Chan, A. H. S., Ashton, H. C., & Krueger, D. A. (2023). *Characterizing manipulation from AI systems*. *arXiv*. <https://arxiv.org/abs/2303.09387> – see also FAccT '23 Companion version: <https://doi.org/10.1145/3617694.3623226>
- Everitt, T., Carey, R., Langlois, E. D., Ortega, P. A., & Legg, S. (2021). Agent incentives: A causal perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13), 11487–11495. <https://doi.org/10.1609/aaai.v35i13.17368>
- Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education*, 9, e46885. <https://doi.org/10.2196/46885>
- Ferreira, J. J., & de Souza Monteiro, M. (2021). The human–AI relationship in decision-making: AI explanation to support people on justifying their decisions. *arXiv*. <https://arxiv.org/abs/2102.05460>
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Insiew, M., & Alfandi, O. (with Al-kfairy, M., Mustafa, D., & Kshetri, N.) (2024). Ethical challenges and solutions of generative AI: An interdisciplinary perspective. *Informatics*, 11(3), 58. <https://doi.org/10.3390/informatics11030058>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Klenk, M. (2024). Ethics of generative AI and manipulation: A design-oriented research agenda. *Ethics and Information Technology*, 26(1), 97. <https://doi.org/10.1007/s10676-024-09745-x>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). <https://doi.org/10.1145/3287560.3287596>
- Raji, I. D., Smart, A., White, R. N., Theron, D., Mitchell, M., ... Gebru, T. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). <https://doi.org/10.1145/3351095.3372873>
- Schemmer, M., Köhl, N., Benz, C., & Satzger, G. (2022). On the influence of explainable AI on automation bias. *arXiv*. <https://arxiv.org/abs/2204.08859>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>

- Wang, Y., Zheng, P., Peng, T., Yang, H., & Zou, J. (2020). Smart additive manufacturing: Current artificial-intelligence-enabled methods and future perspectives. *Science China Technological Sciences*, 63(8), 1600–1611. <https://doi.org/10.1007/s11431-020-1639-9>
- Xie, Q., & Wang, F. (2023). Faithful AI in medicine: A systematic review with large language models and beyond. *medRxiv*. <https://doi.org/10.1101/2023.07.14.23292835>
- Yu, K., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10), 719–731. <https://doi.org/10.1038/s41551-018-0305-z>
- Zhang, G., Chong, L., Kotovsky, K., & Cagan, J. (2023). Trust in an AI versus a human teammate: The effects of teammate identity and performance on human–AI cooperation. *Computers in Human Behavior*, 139, 107536. <https://doi.org/10.1016/j.chb.2023.107536>

Authors' Biographies

- **Dr. Houda BOUMEDIENE** is an Associate Professor at the University of Laghouat, Algeria, specializing in Applied Linguistics. Her scholarly work encompasses critical thinking, Integrating Content and Language in Higher Education (ICLHE), corpus-based analysis, artificial intelligence in education, and adaptive learning methodologies. She has published several academic books and peer-reviewed articles in these fields.
- **Dr. Ouafa OUARNIKI** is an Associate Professor at Ziane Achour University of Djelfa, Algeria, with a research focus on English for Specific Purposes (ESP). She is actively involved in national and international academic forums and has led numerous initiatives in ESP instruction, curriculum development, and teacher professional development. Her contributions extend to academic coordination and leadership roles in international research networks.
- **Prof. Nadežda STOJKOVIĆ** is an Associate Professor at the University of Niš, Serbia, specializing in English for Specific and Academic Purposes. She organizes international ESP conferences and serves as an Advisory Editor for Cambridge Scholars Publishing. Her research focuses on ESP, EAP, academic writing, and applied linguistics.