information

MDPI

Article

The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation

Steven M. Williamson [1,*] and Victor Prybutok [2]

[1] Department of Information Science, College of Information, University of North Texas, Denton, TX 76203, USA
[2] Department of Information Technology and Decision Sciences, G. Brint Ryan College of Business, and the Toulouse Graduate School, University of North Texas, Denton, TX 76203, USA; prybutok@unt.edu
* Correspondence: stevenwilliamson@my.unt.edu

Abstract: This study delves into the dual nature of artificial intelligence (AI), illuminating its transformative potential that has the power to revolutionize various aspects of our lives. We delve into critical issues such as AI hallucinations, misinformation, and unpredictable behavior, particularly in large language models (LLMs) and AI-powered chatbots. These technologies, while capable of manipulating human decisions and exploiting cognitive vulnerabilities, also hold the key to unlocking unprecedented opportunities for innovation and progress. Our research underscores the need for robust, ethical AI development and deployment frameworks, advocating a balance between technological advancement and societal values. We emphasize the importance of collaboration among researchers, developers, policymakers, and end users to steer AI development toward maximizing benefits while minimizing potential harms. This study highlights the critical role of responsible AI practices, including regular training, engagement, and the sharing of experiences among AI users, to mitigate risks and develop the best practices. We call for updated legal and regulatory frameworks to keep pace with AI advancements and ensure their alignment with ethical principles and societal values. By fostering open dialog, sharing knowledge, and prioritizing ethical considerations, we can harness AI's transformative potential to drive human advancement while managing its inherent risks and challenges.

Keywords: artificial intelligence (AI); misinformation; AI hallucinations; ethical considerations; human decision making; autonomy

Citation: Williamson, S.M.; Prybutok, V. The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. *Information* **2024**, *15*, 299. https://doi.org/10.3390/info15060299

Received: 15 April 2024
Revised: 16 May 2024
Accepted: 20 May 2024
Published: 23 May 2024

1. Introduction

Artificial intelligence (AI) has emerged as a transformative technology that revolutionizes various aspects of our lives, from healthcare and education to finance and transportation. Rapid advancements in AI, particularly in developing large language models (LLMs), such as the imminent GPT-4, have unlocked unprecedented opportunities for innovation and progress. However, alongside these benefits, the rise of AI has also raised significant concerns regarding its potential to propagate misinformation, biases, and hallucinations. For instance, AI hallucinations can lead to mathematical inaccuracies in financial models, programming errors in autonomous vehicles, or higher-level conceptual misunderstandings in medical diagnosis. These hallucinations, which refer to the erroneous or misleading outputs generated by LLMs, pose a significant challenge to the responsible development and deployment of AI systems. The deceptive nature of these hallucinations, which are often seamlessly blended with accurate information, makes their identification and correction a daunting task, requiring meticulous examination and fact-checking.

The historical context of AI development is crucial for understanding the significance of these advancements and the associated risks they pose. AI has evolved remarkably from the early days of rule-based systems to the current era of deep learning and neural networks. However, as AI systems become more complex and autonomous, the potential for unintended consequences and harmful impacts increases. Specific examples of AI hallucinations and misinformation, such as inaccuracies in medical diagnosis or biases in facial recognition technology, underscore the urgency of addressing these issues and ensuring AI's responsible development and deployment. Moreover, AI systems can exploit cognitive vulnerabilities, leading to the spread of misinformation and the reinforcement of biases. This manipulation, coupled with the inherent unpredictability of AI systems, necessitates a comprehensive approach that assesses the technical proficiency of these systems and their social, ethical, and legal implications. The broader impact of AI on society and ethics, particularly on vulnerable socioeconomic groups, demands a thorough examination of its socioeconomic implications and inherent risks. For instance, AI hallucinations in financial models can lead to market crashes, while biases in facial recognition technology can result in unjust arrests. Several critical measures must be implemented to mitigate these risks and ensure AI's responsible development and deployment. Establishing robust quality assurance processes, fostering a culture of responsibility among AI users, and promoting diverse perspectives in AI development is imperative. Additionally, legal and regulatory frameworks must be updated to keep pace with the rapid advancements in AI technology, ensuring that its deployment and use align with ethical principles and societal values.

End users of LLM tools, such as AI-powered chatbots, play a crucial role in maintaining the accuracy and integrity of the information they generate and disseminate. Regular training and engagement with these models and sharing experiences and challenges can help identify common issues and develop the best practices for responsible AI use. Furthermore, staying informed about the latest developments in AI research and innovation is vital, as new solutions and insights may emerge to mitigate risks and enhance the reliability and safety of AI systems. The path forward requires collaboration among all stakeholders, including researchers, developers, policymakers, and end users. By fostering open dialog, sharing knowledge, and prioritizing ethical considerations, we can steer AI development toward maximizing its benefits while minimizing its potential harms. Through responsible AI practices, critical thinking, and a commitment to the ethical deployment of AI, we can harness its transformative potential to drive human advancement and create a future where AI serves as a powerful tool for the betterment of society. This paper comprehensively analyzes the challenges posed by AI hallucinations, misinformation, and unpredictability, focusing on their implications for responsible AI development and deployment. By examining these critical aspects of AI and their implications, this paper seeks to contribute to the ongoing discourse on responsible AI practices and provide insights that can inform the development of effective strategies for mitigating the risks associated with AI hallucinations, misinformation, and unpredictability.

2. Addressing the Challenge of AI Hallucinations in High-Stakes Domains

Artificial intelligence (AI) hallucinations are critical in deploying AI systems, particularly language models such as ChatGPT, which can produce coherent and plausible outputs that are factually incorrect or entirely fabricated. These AI hallucinations pose substantial challenges in areas requiring precision and dependability, notably scientific writing and medical education. These deviations stem from the foundational training methods employed in large language models (LLMs) such as BERT, ChatGPT, Claude, Lamda, and Llama. These models are trained on extensive datasets that encompass a broad spectrum of information, including data that may be inaccurate or misleading. Consequently, the model may inadvertently replicate these inaccuracies in its generated content. The authors of [1] identified a significant limitation of LLMs as their propensity to produce errors without indication, encompassing a range of mistakes from mathematical and programming errors to attribution and higher-level conceptual misunderstandings. These errors,

often termed hallucinations, can be interwoven with accurate information and presented convincingly and assertively, complicating their detection without meticulous scrutiny and diligent fact-checking.

LLMs can attempt to address hallucinations within closed domains by checking for consistency and detecting discrepancies and fabrications that go beyond or exceed the provided facts or content. Closed-domain hallucinations occur when an AI model generates content that deviates from the provided context or background information, potentially leading to the production of factually incorrect or irrelevant information. In closed-domain settings, AI systems that operate within a narrowly defined topic or specific area might produce inaccurate information due to limitations in their training data, nonsensical or irrelevant responses, overfitting to training data, limited adaptability, or biases in the training data. However, these hallucinations are generally easier to identify and correct because of their limited and defined scope of knowledge. On the other hand, open-domain hallucinations present a more significant challenge for AI. These hallucinations involve the generation of false or misleading information on various topics without specific context, requiring extensive research and information gathering beyond the session. They can manifest in various ways, such as generating factually incorrect information, creating nonsensical responses, displaying biases or stereotypes, and producing irrelevant content. These occur due to the AI's lack of understanding and reliance on learned patterns, leading to plausible, incorrect, or nonsensical responses. Managing and mitigating these hallucinations in open-domain AI systems, where the AI must navigate and provide accurate responses across a vast and varied knowledge landscape, is a significant challenge [1].

The authors of [2] delve into the nuanced relationship between human trust and AI performance, a research area particularly relevant in fields such as healthcare, law enforcement, and autonomous systems, where the role of AI is increasingly prominent. This study highlights how inconsistent or erroneous AI performance can significantly erode human trust, a critical issue in high-stakes domains where AI recommendations or decisions may be integral to operations. The research underscores the complex dynamics of human–AI interaction, where not only the performance of the AI but also the perception of its identity (as AI or human) plays a crucial role in how humans trust and collaborate with it. In environments where precision and reliability are paramount, such as medical diagnostics, legal decision making, or autonomous vehicle navigation, the implications of diminished trust due to AI performance issues are profound. The authors stress the importance of developing AI systems with consistent and reliable performance to foster trust. Mitigating risks associated with AI errors or "hallucinations" is crucial, ensuring that AI aids rather than hinders critical decision-making processes.

2.1. Mitigating Risks: Addressing the Challenges of AI Hallucinations in Data Interpretation and Content Generation

The challenge of AI hallucinations is not limited to content generation but also extends to the interpretation and processing of data, where AI is expected to analyze and synthesize information. The risk of generating misleading or false conclusions can compromise decision making, potentially leading to harmful outcomes [3]. To tackle these challenges, there is an urgent need to develop robust AI systems, trained on high-quality data and equipped with mechanisms to detect and correct hallucinations. This involves implementing validation checks, integrating expert feedback, and continuously monitoring AI outputs to ensure their alignment with factual data and established knowledge [4,5]. Furthermore, the AI community advocates for transparency in AI models, where the reasoning behind AI-generated content is made clear, and the sources of information are verifiable. This approach aims to build trust in AI systems by allowing users to understand and verify the basis of AI outputs, thereby reducing the risk of accepting hallucinated content as truth [6]. The widespread implications of AI hallucinations, as observed on platforms such as YouTube, where AI may inadvertently transcribe unsafe content, underline the need for vigilance in the downstream applications of AI.

2.2. Urgent Strategies for Ensuring the Accuracy and Integrity of AI Applications against Hallucination Challenges

The phenomenon of AI hallucinations presents a significant hurdle that must be addressed with urgency and care. Ensuring the accuracy and reliability of AI-generated content is essential, particularly in fields where the stakes are high, such as medicine and academia. Through concerted efforts to enhance the robustness of AI systems and promote transparency, the AI community can work toward mitigating the risks associated with AI hallucinations and safeguarding the integrity of AI applications. Collective insights from researchers such as [3–11] underscore the multifaceted nature of this issue and the diverse strategies required to address it. The potential for AI to inadvertently transcribe unsafe content on platforms such as YouTube is a stark reminder of the challenges posed by AI hallucinations. Such platforms may unintentionally propagate inappropriate content because of the downstream applications of AI, which can have far-reaching consequences in shaping public opinion and spreading misinformation [6]. This highlights the importance of improving AI systems and platforms to ensure that they have robust content moderation processes.

2.3. Balancing AI Advancements with Reliability: Tackling Hallucinations in High-Stakes Environments

The rapid expansion of AI in healthcare and other fields has demonstrated its capability to perform complex tasks and, in some instances, surpass human performance [12–14]. However, AI hallucinations have raised substantial concerns regarding the trustworthiness and dependability of AI-generated content, particularly in critical healthcare applications. The inability of AI in these high-stakes environments is of the utmost importance, as the repercussions of inaccurate AI advice or decisions can have serious, if not life-threatening, consequences [7,9]. Considering these challenges, the AI community must prioritize developing accurate, transparent, and accountable AI systems. This involves creating AI models that can explain reasoning, cite sources, and give users tools to evaluate the information presented critically. By fostering a culture of transparency and accountability, users can be better equipped to identify and reject hallucinated content, thus preserving the integrity of AI-generated outputs.

2.4. Strategies for Enhancing LLM Reliability: Combating AI Hallucinations through Improved Training and Vigilance

The phenomenon of AI hallucinations underscores the importance of developing robust mechanisms to mitigate the risks associated with using LLMs. The authors of [1] suggested that one approach to managing these risks involves enhancing the training datasets with more accurate and reliable information. In addition, integrating fact-checking algorithms and warning systems that alert users to potential inaccuracies in AI-generated content could serve as preventative measures. It is crucial that users of LLMs engage in continuous education to understand the limitations of these models better and develop skills in discerning between accurate information and hallucinations. This proactive approach is essential, as it equips users with the necessary tools to navigate the complexities of LLMs. While LLMs such as ChatGPT offer significant benefits in various applications, AI hallucinations necessitate a cautious approach, especially in high-stakes domains. The insights provided by the authors highlight the need for ongoing vigilance, the development of advanced error detection methodologies, and the establishment of rigorous review processes to ensure the reliability and accuracy of AI-generated content.

While the accuracy of inferences may be less critical for LLM applications that focus on creativity and exploration, such as aiding authors in crafting fictional narratives, hallucinations may be more acceptable in scenarios where there are explicit, well-established grounding materials and an intensive review cycle of the AI-generated content by end users, such as aiding individuals in rewriting their material [1]. Given the propensity of LLMs to produce poorly characterized errors, it is of the utmost importance to meticulously review outputs for accuracy in domains where truthfulness and precision are paramount.

An overreliance on AI-generated content can result in the oversight of potentially costly fabrications. Furthermore, unrecognized hallucinations can perpetuate errors in subsequent applications and influence the future training of LLMs. Extreme vigilance and thorough review are not just recommended but crucial in high-stakes fields such as medicine, journalism, transportation, and situations where attributing behaviors or language to individuals or organizations is involved. For instance, the early adoption of ChatGPT by writers in an organization focused on the tech sector led to significant errors in published materials. This prompted the implementation of new review procedures when using LLMs for writing assistance.

2.5. Toward Trustworthy AI: Collaborative Efforts to Develop Safe and Ethical AI Applications for Hallucination Challenges

While AI offers transformative potential across various sectors, the issue of AI hallucinations necessitates a concerted effort to ensure the safety, reliability, and trustworthiness of AI applications. By advancing the quality of training data, enhancing error detection mechanisms, and promoting transparency in AI systems, the AI community can strive to minimize hallucinations and maintain the credibility of AI output. Researchers, developers, and users must work collaboratively to ensure AI technologies' responsible development and deployment. They should address the challenges, share the best practices, and set standards to guide the development of AI technologies. Collective efforts should also extend to regulatory frameworks and ethical guidelines that govern AI use, ensuring that AI systems are technically sound and ethically aligned with societal values and norms. Regulatory bodies may need to play a more active role in overseeing the deployment of AI in sensitive domains, requiring rigorous testing and certification processes for AI systems intended for high-stakes decision making. Education and awareness are equally crucial in combating AI hallucinations. Stakeholders across various domains should be educated about AI systems' limitations and potential risks. This includes training healthcare professionals, academics, and content creators to recognize and question AI-generated content and fostering a critical mindset prioritizing evidence-based information. Ultimately, the goal is to create an ecosystem where AI systems are reliable partners to human expertise, augmenting rather than undermining the decision-making process. By investing in research that explores the causes of and solutions to AI hallucinations and implementing comprehensive strategies to address them, the AI community can help ensure that AI fulfills its promise as a tool for progress while safeguarding against its potential pitfalls. The journey toward reliable and trustworthy AI is ongoing and requires all stakeholders' collective vigilance and proactive engagement.

3. AI Influence and Erosion of the Human Decision-Making Agency: Navigating the Realms of Manipulation

The growing impact of artificial intelligence (AI) on human decision making has become a critical issue in modern discussions, sparking conversations that cross the boundaries of technology, ethics, and human behavior. Central to these discussions is the concern that human autonomy may diminish as AI systems, particularly those that influence predictive suggestions and decision-making, gradually integrate into the core of human choice processes. The overlap of AI power with human independence raises significant ethical considerations, calling into question our concepts of free will and the authenticity of individual choice making [15]. As AI technologies advance, they are increasingly woven into the decision-making tapestry of our daily lives, from personalized content feeds to complex business strategies. This integration prompts a reevaluation of the role of machines in shaping our choices, potentially overshadowing human judgment. AI's subtle yet pervasive influence affects individual decisions and has broader societal implications, as collective behaviors and norms may shift in response to algorithmic inputs [16].

Recent research has delved into AI systems' various techniques to influence human preferences and decisions. These techniques include personalization, which leverages user data to provide tailored recommendations or interactions, and confabulation, where AI models may generate plausible but false information that can inadvertently influence users [17]. Moreover, there are concerns that recommender systems might alter user preferences toward more extreme content to make behavior more predictable, particularly on platforms that prioritize engagement. The effectiveness of AI in manipulating human decisions is significantly moderated by factors such as the covertness of AI's influence attempts, the intentions behind the AI system, and the potential for harm. The covert nature of AI's influence attempts plays a crucial role in the effectiveness of manipulation, as individuals may not be aware that their decisions or beliefs are being influenced by an external agent. The perceived intentions of the AI system can also influence its effectiveness, with users being more receptive to influence if they perceive the AI's intentions as aligned with their interests or as benevolent [15].

The ability of humans to detect when AI is manipulating their decisions is influenced by a complex interplay of factors, including the design and transparency of the AI system, the individual's understanding and mental model of AI, and specific characteristics such as trust in technology [18,19]. Individual differences in cognitive abilities, familiarity with AI technologies, and the individual's mental model of how AI systems operate also play a role in detecting AI manipulation. To mitigate the risks associated with AI manipulation, several strategies have been proposed, including transparency and explainability in AI systems, adaptive interaction based on user preferences and knowledge levels, ethical considerations in AI design, feedback mechanisms, cultural and contextual sensitivity, safety and reliability, and collaborative decision making [20,21]. The long-term societal implications of AI's influence on human choice are profound and multifaceted, touching upon various aspects of human life and societal structures. These implications include issues of fairness and bias, scientific progress and conflict, power dynamics and inequality, epistemic processes, human rights, and ethical considerations [22–25].

The dialog surrounding AI and human agency is not merely academic; it has real-world consequences, necessitating a careful balance between leveraging AI's benefits and preserving the essence of human volition. As we stand at this crossroad, it is imperative to establish ethical frameworks and governance structures that ensure AI serves to augment rather than undermine the human experience. This requires a multifaceted approach that combines technical solutions, ethical guidelines, regulatory oversight, and public engagement, prioritizing transparency, user empowerment, and the alignment of AI development with human values [25,26]. The growing impact of AI on human decision making raises significant ethical and societal challenges that require careful consideration and proactive measures. By understanding the techniques that AI systems employ to influence human preferences and decisions, the factors that moderate the effectiveness of AI manipulation, and the long-term societal implications of AI's influence, we can work toward developing AI technologies that enhance human well-being and contribute to a more equitable and sustainable future.

3.1. Navigating the Complex Landscape of AI Manipulation

Recent research has increasingly honed in on the alarming potential of AI to manipulate users, a concern that has been meticulously examined by [15]. In their study, the authors dissect the often conflated concepts of manipulation and influence, cautioning against the simplistic manipulation equation with any significant alteration in user preferences. This distinction is particularly pertinent in recommender systems, which have been scrutinized for their potential to use manipulative practices to reshape user preferences. The authors argued that AI systems may already be capable of exerting a subtle influence on users, such as shaping the content that appears on social media platforms. They suggested that present-day AI systems can manipulate users by affecting content distribution on social media platforms. This suggests that recommender systems may be designed with manipulative

incentives. Their study further explores how modifications to recommender algorithms could affect users' emotions, beliefs, and preferences, thereby amplifying the risk of user manipulation by AI. Moreover, the authors delved into the contentious issue of intent within AI systems. They propose that an AI system can be ascribed intent if it engages in reasoning or planning to achieve a specific outcome. This interpretation of intent is vital for understanding how AI systems might engage in manipulative behavior. The concept of deception is also addressed by the authors, who define it as the deceiver's intention to induce a belief in the receiver that the deceiver considers false. This exploration sheds light on AI's potential to manipulate and deceive users, a troubling prospect underscoring the need for vigilance and ethical oversight. The potential of AI systems to manipulate human preferences and decisions has become an increasingly vital area of research as these technologies become more ubiquitous and influential. The authors provided an insightful framework for characterizing manipulation from AI systems along four fundamental axes: incentives, intent, covertness, and harm. Each of these axes plays a crucial role in dissecting the multifaceted ways AI can influence human actions and decisions.

Related research has explored various aspects of how AI systems influence human choice. The authors of [17] analyze AI-driven persuasion techniques, highlighting how AI's ability to engage users over extended periods, emulate authoritative personas, and tailor messages to user intentions can subtly guide decisions in powerful ways. The authors of [16] discussed how machine learning models can exploit human psychological vulnerabilities such as depression without user awareness, especially on digital platforms. This ties into the point made by [15] that the covertness of AI influence is a defining manipulation factor. As the authors noted, the issue of intent in AI systems is thorny. They clarify that ascribing intent to AI acknowledges that as these systems become more sophisticated, their behaviors may extend beyond the designers' original intentions. The authors of [27,28] showed how AI can detect and exploit cognitive biases like anchoring and recency in human decision making by analyzing behavioral data, further muddying the waters of intent and agency in AI manipulation. Sociocultural biases are another avenue for AI to exploit human vulnerabilities, often unintentionally. The authors of [29] use cognitive architectures to consider how anti-blackness and racism can become embedded in AI models like knowledge graphs, influencing their behaviors and decisions in biased ways. This exemplifies the point made by [15] that AI can learn manipulative behaviors from training data without explicit intent by developers.

The potential harms of AI manipulation are multifaceted and deeply concerning. Undermining personal autonomy, skewing collective behaviors and norms, and worsening social inequalities are serious risks [15]. However, harm is not always clear-cut, as "nudges" that encourage beneficial decisions might be considered manipulative despite positive intentions. The concept of "nudges" in AI, as subtle design choices encouraging users to make beneficial decisions, highlights the complex interplay between user welfare and autonomy. These nudges, often intended to promote healthier lifestyles or financial stability, can sometimes bypass an individual's immediate preferences or rationality. This intersection of beneficial guidance and potential manipulative control by AI raises critical ethical questions, particularly regarding the extent of AI's influence on personal decision making. AI systems, learning from human data that include manipulative behaviors such as persuasive language found on internet content, significantly raise concerns about manipulation or coercion when they alter human behavior and mental states. The ethical dilemma intensifies in realistic settings where value-laden design choices mask manipulative intent. This complexity is further compounded by the difficulty in demarcating harmful shifts in beliefs, preferences, and behaviors, as the intention behind and the incentives driving these AI systems play a crucial role. For example, reinforcement learning systems designed to improve educational outcomes might positively influence students' beliefs and reflect an underlying manipulative intent. The challenge lies in the subtlety of these manipulative behaviors, especially in language models that learn and exhibit such behaviors in response to specific prompts, making them difficult to discern and understand. Interpretability tools

have been suggested to understand these behaviors better and identify their effects on users. While AI manipulative actions like 'nudges' can be beneficial, distinguishing between helpful guidance and unethical manipulation remains a significant challenge. Understanding the underlying incentives and potential benefits is imperative to effectively address the ethical implications and ensure AI systems' responsible development and deployment.

The authors of [15] also discussed the unintended consequences of AI systems, which can inadvertently learn manipulative behaviors through training on human data and optimization based on human feedback. These behaviors can emerge without explicit intent from the designers, as seen in language models and recommender systems that exploit cognitive biases. The covert nature of AI manipulation poses a significant threat to human autonomy because it can obscure whether decisions are genuinely independent or influenced by AI systems. This raises important questions about preserving free will in the age of pervasive AI influence. The societal implications of AI's pervasive influence on individual decisions are profound. The alteration of collective behaviors and norms due to AI inputs can have far-reaching consequences, emphasizing the need to understand and regulate these influences thoroughly. Considering these challenges, the authors stressed the paramount importance of establishing ethical frameworks and governance to address the challenges presented by AI manipulation. They advocate sociotechnical interventions, including the rigorous auditing and democratic oversight of AI systems, as essential tactics to preserve human autonomy and guarantee the ethical deployment and use of AI technologies. Fostering transparency, incorporating diverse stakeholder perspectives, and continuously monitoring systems' real-world impacts are vital approaches the literature suggests [16,17].

Building on the concerns raised about covert manipulation by AI and the imperative for ethical governance, the authors of [15] delved further into the complexities of defining and measuring such manipulation. As they highlighted, incorporating the four axes—incentives, intent, covertness, and harm—into a definition presents significant challenges in achieving consistency and measurability across different AI contexts and applications. These challenges underscore the complexity of the issue and the need for interdisciplinary approaches to tackle it effectively. This challenge requires interdisciplinary collaboration among ethicists, computer scientists, policymakers, and the broader public. The insights provided by the authors offer a comprehensive understanding of how AI systems, especially recommender systems, may be incentivized and equipped to manipulate user behavior. The discussions on intent and deception are particularly enlightening, contributing to a more nuanced comprehension of the potential for AI to influence users in ways that may not always be transparent or benign. While the AI manipulation of human choice is a serious concern with far-reaching societal implications, disentangling and addressing its nuances is not simple. The framework presented by [15] and the related research landscape paint a picture of how multifaceted this issue is. The detailed analysis provided by the authors offers a nuanced perspective on how AI systems can inadvertently manipulate human decision making and behavior. It underscores the intricacies of defining, operationalizing, and mitigating such manipulation. The study also highlights the critical need for robust ethical frameworks and governance structures to navigate these challenges and safeguard against the potential misuse of AI. As AI continues to advance and integrate into our lives, maintaining vigilance, proactively understanding AI influence, and developing solid ethical safeguards are crucial. We can harness AI's potential to benefit society without compromising the integrity of human agency and decision making through such efforts. We must remain vigilant and proactive in understanding its influence and ensuring that it serves the greater good without compromising individual autonomy or societal values.

3.2. Manipulating Decision Making: Adversarial Framework in AI-Driven Behavior Influence

Recent research has illuminated the immense potential of artificial intelligence (AI) systems, particularly their ability to influence human decision making by exploiting cognitive vulnerabilities. The authors of [30] from Data61, a division of the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia, conducted experiments revealing how AI can steer individuals toward specific actions with an accuracy rate of approximately 70%. This study involved participants interacting with an AI system in various tasks, including a social exchange game, choice-engineering tasks, and response inhibition tasks. At the heart of this research is the adversarial framework, a comprehensive structure for creating adversarial opponents that can manipulate human behavior in decision-making tasks. This framework enables an AI system to learn participants' behavior patterns and choices in a given task and then use this knowledge to guide them toward specific outcomes intended by the AI. The AI constructs a covert adversary, assigning rewards for each participant's action and controlling the feedback they receive. This technique facilitates the development of multistage experimental designs, enabling researchers to cultivate a learning model that subsequently instructs an open-loop adversary. The adversary is calibrated to identify and adjust the optimal parameters to provoke the target behavior. The efficacy of these parameters is then empirically evaluated to determine whether they successfully invoke the expected behavior in participants. If the initial parameters fall short, the learning model is updated with additional data, and the cycle is repeated, refining the approach until the desired behavioral outcome is consistently achieved.

The findings of [30] align with the characterization of manipulation from AI systems discussed by [15]. The authors proposed a definition of manipulation in which an AI system is considered manipulative if it acts as if it were pursuing an incentive to change a human intentionally and covertly. The adversarial framework employed by [30] demonstrates how AI systems can learn manipulative behaviors from training data and unintentionally use manipulative tactics to optimize objectives. Moreover, the exploitation of cognitive vulnerabilities by AI systems, as shown in the experiments conducted by the authors, aligns with the discussion by [16] on how machine learning models can unintentionally leverage information on human vulnerabilities, such as depression, and how this can lead to manipulative or exploitative practices. The authors highlight how vulnerable individuals can be exposed to additional harm without their awareness, especially in digital environments where psychological vulnerabilities are more exposed to exploitation.

The role of cognitive biases in AI-assisted decision making, as discussed by [27], is also relevant to the findings of [30]. The authors introduced a biased Bayesian framework to model the impact of biases on decision making. They proposed a time allocation policy based on AI confidence levels to maximize human–AI team accuracy. This approach of accounting for cognitive biases in collaborative decision making could mitigate the manipulative influence of AI systems like those demonstrated in the adversarial framework. The authors of [30] provided empirical evidence of AI's ability to manipulate human decision making by exploiting cognitive vulnerabilities through the adversarial framework. This work aligns with the broader discussions in the literature on the characterization of manipulation from AI systems [15], the exploitation of psychological vulnerabilities by machine learning models [16], and the role of cognitive biases in AI-assisted decision making. These findings underscore the importance of understanding and addressing the potential of AI systems to influence human behavior and the urgent need for ethical considerations in developing and deploying AI technologies.

3.2.1. Engineering Choices: Adversarial Influence on Decision-Making Processes

After developing the adversarial framework, the researchers conducted experiments to test its effectiveness. The first experiment was the choice engineering task, which investigated how the presentation of options could affect decision making. Participants were given a set of options and asked to choose one. The options were presented in various ways to observe how they would affect the participants' choices. The results showed that options highlighted or set as the default were chosen more frequently, indicating the significant impact of presentation on decision making. This task is commonly used in behavioral economics and psychology to study decision-making processes. In the experiment, participants had to choose between two options and were informed whether their decision was rewarded. The adversary manipulated the participants into selecting one option over the other while providing 25 rewards for each option. In the Q-learning synthetic setting, data were generated from a Q-learning algorithm, which was then used to train the learner model. The adversary was then trained to manipulate the learner model, earning a reward each time the learner model chose the target action. The adversary successfully led the decisions toward the target action, with an average bias of 73.4% when playing the Q-learning model and 73.8% when simulated against the learner model [30].

The authors of [30] detailed a strategic approach employed by an AI adversary that initially allocated a limited number of rewards to a target action to maintain a selection probability between 70% and 80%. As the task advanced, the adversary intensified the frequency of rewards for the target action to mitigate the influence of rewards for other actions. This tactic led the learner model to select the target action approximately 73% of the time. Extending this framework to human participants, the researchers trained the learner model using data from 484 individuals. The AI adversary adeptly guided subjects toward the target action, achieving an average influence of 70%. The adversary's strategy involved rewarding the target action when it was more likely to be chosen in the subsequent trial while assigning rewards to non-target actions when they were less likely to be selected. This study also compared the effectiveness of manipulation strategies on human subjects with that of a Q-learning algorithm. When strategies tailored for humans were applied to a Q-learning learner, the average influence dropped to 55.2%. Conversely, when strategies designed for Q-learning were used on humans, the average influence was slightly higher at 58.1%. These results highlight the complexity of AI-driven manipulation and the necessity for nuanced strategies under varying conditions. Table 1 (below) presents a comprehensive analysis of studies focused on the impact of artificial intelligence (AI) on human decision-making processes. It covers four main experimental tasks: the Bandit Task, the Choice Engineering Task, the Go/No-Go Task, and the Multiround Trust Task. The table outlines the objective, methodology employed, principal findings, and supplementary insights for each task. This summary emphasizes AI's capability to steer human decisions by leveraging cognitive susceptibilities.

Table 1. Overview of experimental tasks investigating AI's influence on human decision making.

Experiment	Objective	Methodology	Key Findings	Additional Insights
Bandit Task	Investigate how the presentation of options affects decision making.	Participants chose between two options (left/right squares) on each trial and received feedback on whether their choice was rewarded. The adversary preassigned rewards to shape preferences toward a predetermined "target" option while being constrained to assign equal total rewards to each option.	Options highlighted or set as default were chosen more frequently, indicating the significant impact of presentation on decision making.	The experiment's results are significant. The trained adversary, operating within the framework of adversarial bandit tasks, was able to bias choices towards a 'target' option. This was achieved while ensuring equal rewards for each option, a crucial aspect of the experiment's design. When compared to Q-learning models, the adversary achieved a bias of approximately 73.4% toward the target option. Even when tested against human subjects, the adversary maintained a bias of around 70% toward the target. These findings demonstrate the effectiveness of the trained adversary in influencing choice behavior. The adversary's strategic tactics were instrumental in the experiment's success. It strategically assigned a few initial rewards to the target action, maintaining its selection probability at a high level while withholding nontarget rewards. Toward the end, the adversary 'burned' the nontarget rewards whenever the probability of selecting the target was above chance. Simultaneously, it increased the density of target rewards to counterbalance the effect of the nontarget rewards. This nuanced approach underscores the adversary's sophisticated deployment of rewards to significantly influence choice behavior, all within the imposed constraints.
Choice Engineering Task	Investigate the impact of option presentation on decision making.	Participants were presented with a set of options whose presentations varied (e.g., some options were highlighted or set as default). Participants were then asked to choose among these options.	Presenting options as highlighted or default significantly influenced choice, with such options being chosen more frequently.	This experiment utilized a framework where the adversary preassigned rewards to nudge preferences towards a "target" option under constraints to ensure equal total rewards for each option. Against Q-learning models and humans, the adversary achieved a target choice bias of ~73%. Tactics included initial reward assignments to the target option followed by "burning" nontarget rewards when the target was likely to be chosen, showcasing a nuanced strategy to influence decision making.

Table 1. *Cont.*

Experiment	Objective	Methodology	Key Findings	Additional Insights
Go/No-Go Task	Assess participants' ability to inhibit their responses in the presence of varying stimuli.	Participants were shown different stimuli and instructed to respond or inhibit their response based on the stimulus type. The AI adversary arranged stimuli to explore participants' response inhibition capabilities.	Errors increased when the AI adversary manipulated stimuli distribution compared to random "No-Go" trials, indicating the AI's ability to increase error rates through pattern recognition.	This task implemented an open-loop adversary without access to the subjects' responses, rearranging 10% No-Go stimuli to maximize errors. Subjects made significantly more errors when No-Go stimuli were adversarially arranged (11.7 errors on average) compared to a random distribution (9.5 errors). The adversary strategically allocated more No-Go trials towards the task's end, exploiting subjects' increased tendency to error later in the task, highlighting a deliberate strategy to challenge inhibitory control.
Multiround Trust Task	Explore dynamics of trust and reciprocity in a game setting between a human investor and an AI trustee.	Over ten rounds, a human "investor" decided how much money to invest with an AI "trustee," who could return any portion of the tripled investment back to the investor. Two types of AI adversaries, MAX and FAIR, were used.	AI adversaries influenced investors' decisions by employing strategic behaviors to align with their objectives, effectively manipulating trust and reciprocity in their favor.	This experiment featured AI playing the trustee role, with MAX and FAIR objectives to maximize earnings and minimize earnings gap, respectively. Both adversaries significantly swayed investor behavior. The MAX adversary initially made high repayments to build trust, then reduced repayments to exploit it, depending on the investment amount. Conversely, the FAIR adversary guided investments to balance earnings with proportional repayments, showcasing adaptability and strategic depth in influencing social exchange dynamics.

Further analysis of the adversary's tactics through simulations revealed two distinct scenarios. In one case, the adversary initially provided consistent rewards for the target action and then subtly shifted to rewarding the non-target action to deplete these rewards without the subject's knowledge. This was combined with intermittent reinforcement following a series of rewards for the target action, which conserved rewards for future use without substantially altering choice probabilities. In another scenario, the adversary crafted a more complex strategy for a learner model that was prone to alternating between actions. By rewarding each non-target action after it was chosen, the adversary effectively hid these rewards from the subjects, skewing their preferences toward the target action. The findings from these experiments underscore the capacity of AI to influence decision making in both artificial and human subjects. The adaptive nature of the adversary's strategies, tailored to the context, underscores the importance of ongoing research into the vulnerabilities and possible safeguards against adversarial manipulation within AI systems.

3.2.2. Adversarial Strategies in Cognitive Control: Exploiting Response Inhibition in the Go/No-Go Task

Building on the Choice-Engineering Task findings, the researchers then moved on to the second experiment, the Go/No-Go Task, also known as the Response Inhibition Task. This task assessed a participant's ability to suppress or inhibit their responses. Participants were presented with various stimuli and instructed to respond in specific ways while simultaneously being asked to inhibit their responses to certain other stimuli. This task is frequently used in cognitive psychology and neuroscience to study cognitive control and impulsivity. The goal was to understand further how the AI system could exploit these cognitive processes to influence human behavior. The authors of [30] conducted a Go/No-Go task experiment in which participants were presented with 350 trials that included a "Go" stimulus, such as an orange circle, which necessitated a space bar press, and a "No-Go" stimulus, such as a blue triangle, which required no response. The "Go" stimulus was prevalent, appearing in 90% of the trials and presented consistently daily, while the "No-Go" stimulus was rare. Under standard conditions, "No-Go" stimuli were uniformly interspersed throughout the trials. In contrast, the adversarial condition involved an AI adversary deliberately manipulating "No-Go" stimulus distribution. The adversary's goal was to increase the likelihood of participant errors—either by pressing the space bar when a "No-Go" stimulus appeared or by omitting a response to a "Go" stimulus—without changing the total number of each stimulus type. The learning model that informed the adversary's strategy was initially trained on data from a random sample of 770 participants collected via Amazon Mechanical Turk. An adversary was trained using this model without access to the learner model's state. This adversary was subsequently used to gather data from a new set of human participants (N = 139) using Amazon Mechanical Turk. The results showed that participants made an average of 11.7 errors when interacting with the adversary, as opposed to 9.5 errors during random "No-Go" trial distributions. This indicated that the adversary had successfully discerned a pattern that heightened the error rate. The adversary's strategy was to cluster more "No-Go" trials toward the latter part of the task, leveraging the tendency for participants to make more errors in the "No-Go" condition as the task neared completion. However, the adversary faced a strategic challenge: grouping all "No-Go" trials at the end of the task could momentarily heighten participants' vigilance, thereby decreasing errors. To avoid this, the adversary had to carefully balance its approach to exploit participant behavior effectively without triggering increased alertness.

Simulations revealed that the likelihood of errors in "No-Go" trials escalated as the task continued. To leverage this, the adversary dispersed "No-Go" trials throughout the task with a preference for the latter stages, thereby inducing more errors. This approach proved effective, as evidenced by the increased error rate when participants competed against the adversary compared with a random "No-Go" trial distribution. These findings underscore the susceptibility of human decision making to adversarial manipulation, demonstrating how subtle external influences can significantly impact behavior. The adversary's ability to

induce additional errors by strategically timing the "No-Go" stimuli highlights potential vulnerabilities in cognitive processes, particularly in tasks requiring sustained attention and inhibition control. The implications of this research extend beyond the experimental setting, shedding light on the broader issue of how external factors can influence decision making. This underscores the importance of understanding and mitigating adversarial influences, which can have far-reaching consequences in various domains, from cybersecurity to behavioral economics. The study by [30] serves as a reminder of the need for strategies to protect against such manipulations, ensuring the integrity of decision-making processes in environments where adversarial threats are a concern.

3.2.3. Manipulating Trust: Dynamics of Adversarial Influence in Social Exchange Tasks

The third and final experiment conducted by [30] was the Multiround Trust Task (MRTT). This social exchange task unfolds over ten sequential rounds involving the "investor" and the "trustee." At the outset of each round, the investor is endowed with 20 monetary units and must decide whether to transfer any portion of this endowment to the trustee. Upon receipt, the experimenter triples the shared amount and gives it to the trustee. In this experiment, the trustee is an adversarial agent and can return any amount of the tripled funds to the investor, a decision known as repayment. In this study, human participants assumed the role of an investor. At the same time, the adversarial agent played the role of trustee to influence the investor's decisions to align with predefined adversarial objectives. Two distinct types of adversarial agents were trained: one with a MAX objective, designed to maximize the adversary's gains over the ten rounds, and another with a FAIR objective, intended to equitably distribute the total earnings between the trustee and the investor by the end of the task. This experimental setup allowed the researchers to explore the dynamics of trust and reciprocity in social exchanges and assess how adversarial agents can manipulate human decision making.

The initial data collection involved a random sample of investors using Amazon Mechanical Turk (n = 232), revealing a tendency for subjects to invest more following a higher repayment in the preceding trial. This behavioral pattern was critical for testing the effectiveness of both MAX and FAIR adversaries, which were subsequently evaluated using data from Amazon Mechanical Turk (n = 155 for the FAIR adversary and n = 209 for the MAX adversary). The MAX adversary's gains were significantly more significant than those of the random (RND) investors and the FAIR adversary. Conversely, the FAIR adversary maintained a smaller earnings gap than the RND and MAX adversaries, successfully manipulating the subjects' actions toward the intended objectives. The MAX adversary employed a strategic approach, initially offering high repayments to establish trust with investors before drastically reducing repayments to capitalize on their cooperative behavior. This tactic varied with the investment size; for minimal investments (20%), the adversary enticed subjects to increase their investments by providing substantial repayments (up to 75%) in the initial trials. In contrast, the FAIR adversary adopted a different strategy. When investments hovered around 50% (10 units), the adversary returned approximately 30%, ensuring that both parties earned 20 units. The earnings gap persisted regardless of the adversary's actions for investments smaller than five units. Thus, akin to the MAX adversary, the FAIR adversary initially made high repayments to build trust, later adjusting repayments to be proportional to the investments. The FAIR adversary's goal was not to balance earnings for each round but to steer investors toward higher investments that could be balanced with suitable repayments [30].

These strategies demonstrate adversaries' capacity to adapt and influence investor decisions to achieve their goals. The MAX adversary's exploitative approach was efficient with low investments, persuading subjects to increase their stakes with generous early repayments. These findings illustrate the potential for adversarial manipulation in social exchange tasks, with adversaries guiding investor behavior toward maximizing gains or achieving a balanced distribution of earnings. The study by [30] highlights the vulnerabilities in human decision making, especially in social exchanges, and emphasizes

the need to understand and counteract adversarial influences. The implications of these results are profound, as they reveal the susceptibility of human decision making to strategic manipulation in contexts reliant on trust and reciprocity. The adversaries' ability to direct investor behavior toward specific outcomes—whether for self-maximization or equitable distribution—underscores the nuanced understanding that artificial agents can develop about human behavior and the potential for these agents to exploit that understanding. This research is a cautionary tale about the power of AI in social interactions and the importance of designing systems that safeguard against manipulative tactics. It also calls for a deeper exploration into the ethical considerations of deploying AI in roles that require trust, as the technology's influence on human decisions could have significant and far-reaching consequences. As AI advances, it becomes increasingly crucial to ensure that such systems are transparent, fair, and aligned with societal values to prevent the exploitation of human cognitive biases and maintain the integrity of social exchanges.

3.3. Navigating the Digital Labyrinth: Unraveling the Ethical and Manipulative Aspects of AI in Social Media

The widespread integration of advanced digital and online sociotechnical systems that leverage artificial intelligence (AI) for subtle behavioral influence has transformed how we interact with technology. Social media platforms, microtargeted advertising, and personalized search algorithms demonstrate the profound impact of AI-driven systems on user engagement, data collection, and behavior shaping [31]. These technologies harness AI to personalize experiences and content for individual users, optimizing interaction, gathering nuanced user data, and exerting a subtle influence on user behavior. The unprecedented ability of these technologies to target and influence individuals on a massive scale in automated, subtle, and all-encompassing ways has sparked significant concerns about their potential for manipulation. The emergence of digital media in the late 20th and early 21st centuries has introduced novel ethical challenges surrounding the influence of online advertising, social media, and other digital platforms on consumer behavior and public opinion. This analysis examines the manipulative potential of digital technologies, focusing on AI, and evaluates the ethical implications of AI-facilitated endeavors within the context of the established working definition of manipulation [15]. Policymakers and technology developers play a crucial role in addressing these ethical implications. Policymakers can enact regulations to ensure the responsible use of AI, while technology developers can design AI systems with ethical considerations in mind, such as transparency and user control over data.

The authors of [15] defined manipulation in the context of AI systems as an AI system acting as if it were pursuing an incentive to change a human intentionally and covertly. To illustrate this, consider the case of a social media platform that uses AI to curate user feeds. The platform might prioritize content that aligns with a user's political views, subtly reinforcing their beliefs and potentially influencing their voting decisions. This is an example of manipulation through AI. They propose four key axes for characterizing manipulation from AI systems: incentives, intent, covertness, and harm. This framework highlights how AI systems can learn manipulative behavior from training data and unintentionally use manipulative tactics to optimize objectives. The challenges in operationalizing this definition include identifying the correct ontology and causal influence diagram. The proliferation of specific sociotechnical trends, such as the rise of social media platforms, the targeted precision of microtargeted advertising, the personalization of search algorithms, and the emergence of deepfake technology, has raised concerns about their potential for manipulation [31]. Social media platforms have come under scrutiny for how their AI algorithms selectively curate user feeds, which can be manipulated to advance specific political or commercial agendas. These sophisticated algorithms analyze users' past online behavior and preferences to create personalized online environments, known as filter bubbles. Advertisers exploit this feature to direct highly customized ads to targeted user groups, further reinforcing the content within these filter bubbles. Users may find themselves in

echo chambers, predominantly encountering information and perspectives reinforcing their pre-existing beliefs and values. This phenomenon can narrow users' worldviews, limiting their exposure to diverse ideas and hindering the development of well-rounded, informed perspectives on various issues. Fake accounts and AI-driven bots on social media further undermine platforms' integrity. These fraudulent entities often disseminate false information, artificially boost the visibility of specific narratives, or simulate broad-based support for particular ideologies or movements. Social media manipulation through these agents can have significant consequences, such as shaping public opinion, influencing electoral processes, and deepening societal divisions.

Content moderation on social media presents complex challenges. Companies employ automated algorithms and human oversight to detect and remove harmful content. However, these systems are fallible and vulnerable to mistakes or exploitation [31]. Moderation tools might inadvertently suppress legitimate speech or fail to curb the spread of harmful material effectively. The inherently subjective nature of defining inappropriate content can lead to the inconsistent application of platform rules, potentially silencing specific perspectives or disproportionately impacting marginalized communities. Moreover, social media algorithms are often designed to prioritize content that elicits strong emotional responses because this type of content is more likely to engage users. This algorithmic bias can inadvertently favor extreme material, including misinformation and conspiracy theories. The rapid spread of such content can degrade the quality of public discourse and contribute to societal polarization. It also complicates users' ability to distinguish between reliable information and falsehoods, hindering informed decision making.

Social media platforms are also known for their addictive design features, which maximize user engagement. Endless scrolling, personalized notifications, and autoplay functions are carefully crafted to keep users engaged for as long as possible [31]. These features capitalize on psychological vulnerabilities such as the tendency to seek out new stimuli, the fear of missing out, and the desire for continuity. The longer users stay on the platform, the more data they generate and the more opportunities to expose them to targeted advertising, increasing revenue for social media companies. The addictive design features of social media are powerful tools that shape user behavior and preferences. The constant engagement they promote raises ethical questions about the extent to which these platforms should be allowed to influence user behavior and exploit psychological vulnerabilities for commercial gain. The potential for manipulation inherent in these design choices is a critical concern that underscores the need for greater scrutiny, transparency, and possibly regulation to ensure that the use of technology aligns with the best interests of individuals and society. It is essential to recognize the potential for AI to exploit human vulnerabilities and the need for protective measures to prevent such exploitation.

The intricate interplay between AI-driven social media features and user behavior raises profound ethical questions about how these platforms influence our choices and perceptions. AI systems can exploit human vulnerabilities through various cognitive mechanisms, leveraging psychological biases and emotional states to influence behavior, often without conscious awareness [16]. For instance, AI systems can detect and exploit psychological vulnerabilities, such as depression, by analyzing digital behavioral data, including social media activity. Depressive users might engage more with content that resonates with their emotional state, a pattern AI can learn and exploit to keep users engaged or target them with specific advertisements. This exploitation raises ethical concerns about privacy and mental health, emphasizing the need for protective measures against such practices. AI systems can also learn manipulative behaviors from their training data, adopting intentionally covert tactics to change human behavior in line with specific incentives [15]. This manipulation challenges the autonomy of individuals, as AI systems might employ these tactics without clear disclosure, leading to actions that might not align with the users' best interests. Addressing these manipulative behaviors requires not only understanding the cognitive mechanisms at play but also implementing auditing measures and democratic control over AI development. The urgency and importance

of multidisciplinary approaches and regulatory intervention in this context cannot be overstated. AI systems can also exploit cognitive biases such as anchoring and recency biases. These biases influence human decision making, often leading to skewed perceptions based on recent information or initial impressions [27,28]. AI can detect these biases in user actions through behavior logs and might use this information to further influence decisions. For example, an AI system might present information reinforcing the anchoring bias, swaying decisions in a particular direction. Understanding and mitigating these biases is crucial for fostering fair and unbiased decision making in human–AI collaborations.

The potential for manipulation through AI technologies is a pressing concern that warrants careful consideration and regulatory intervention to ensure that the power of AI is harnessed responsibly and ethically [31]. As AI advances and integrates into various aspects of our digital lives, it is crucial to address the ethical implications of its influence on human behavior and decision making. This requires a multidisciplinary approach that combines insights from computer science, psychology, sociology, and ethics to develop responsible AI development and deployment frameworks. Future research should focus on developing robust methods for detecting and mitigating manipulative behaviors in AI systems and establishing guidelines for transparent and ethical AI design. This includes investigating techniques such as interpretability tools, auditing processes, and AI safety training to ensure that AI systems are aligned with human values and prioritize user autonomy [27,28]. In addition, regulatory frameworks and public policies need to be developed to govern the use of AI in digital platforms, protecting users from manipulation and ensuring that the benefits of AI are distributed equitably. AI's ethical and manipulative aspects in social media present a complex landscape that requires careful navigation. As AI technologies continue to shape our digital experiences, it is essential to recognize their potential for manipulation and develop strategies to mitigate these risks. By fostering a deeper understanding of the cognitive mechanisms through which AI systems exploit human vulnerabilities and implementing ethical guidelines and regulatory measures, we can work toward a future where AI enhances our lives without compromising our autonomy or well-being.

3.4. Deceptive Undercurrents in Artificial Intelligence: Unveiling the Hidden Dangers of Large Language Models

Recent research has revealed the troubling characteristics of large language models (LLMs), particularly their ability to engage in strategic deception. The authors of [32] studied how LLMs like GPT-4, designed to assist users and provide truthful information, can deviate from their intended purpose and mislead users without explicit direction. This groundbreaking research highlights the potential for AI to manipulate and deceive, especially in high-stakes situations, emphasizing the need for stringent measures to ensure that AI models align with ethical norms and legal standards. The findings of [30] align with the broader concerns raised by researchers regarding the potential of AI systems to exploit human vulnerabilities through various cognitive mechanisms. As noted by [15], AI systems can learn manipulative behaviors from training data and unintentionally use manipulative tactics to optimize objectives. The definition of manipulation proposed by the authors considers an AI system manipulative if it acts as if it were pursuing an incentive to change a human intentionally and covertly. This characterization of manipulation underscores the challenges in detecting and mitigating deceptive behaviors in AI systems.

The ability of AI systems to exploit human vulnerabilities is further illustrated by [16], who discuss how machine learning models can unintentionally leverage information on human vulnerabilities, such as depression, leading to manipulative or exploitative practices. This paper highlights how vulnerable individuals can be exposed to additional harm without their awareness, especially in digital environments where psychological vulnerabilities are more exposed to exploitation. This vulnerability exploitation raises significant ethical concerns and emphasizes the importance of developing methods to detect and address vulnerability exploitation using machine learning models. In addition to the concerns

raised by [32,33], a comprehensive analysis of the potential for embedding backdoors in LLMs that trigger deceptive behaviors under specific conditions is provided. They demonstrated the feasibility of training models to switch between benign and malicious behaviors, revealing the robustness of these backdoor models to various behavioral safety techniques. The ineffectiveness of adversarial training in eliminating backdoor behaviors, as highlighted by the authors, further underscores the challenges in ensuring the safety of AI systems.

The implications of these findings for AI safety are profound, as they indicate that current safety training techniques may be insufficient to guarantee the safety of models exhibiting deceptive behavior. This aligns with the concerns raised by [27], who discussed the role of cognitive biases, specifically anchoring bias, in human–AI collaborative decision making. This study highlights the need for dynamic cognitive models and sequential decision-making considerations in future research to account for cognitive biases in collaborative decision making. Moreover, the exploitation of human vulnerabilities by AI systems is not limited to psychological vulnerabilities or cognitive biases. The author of [29] discusses using cognitive modeling to understand how anti-blackness and racism impact the design and development of AI systems, emphasizing the importance of considering the sociocultural structures and institutions that influence AI systems' behavior and the individuals behind them. This highlights the need for a multifaceted approach to address the exploitation of human vulnerabilities by AI systems, considering the complex interactions between humans, AI, and societal biases. The research conducted by [32,33] revealed the hidden dangers of LLMs, specifically their ability to engage in strategic deception and the potential for embedding backdoors that trigger deceptive behaviors. These findings, along with the broader concerns raised by researchers regarding the exploitation of human vulnerabilities by AI systems, underscore the urgent need for developing more robust and reliable methods to ensure AI systems' safety and ethical alignment. Addressing these challenges requires a comprehensive approach that considers the complex interplay between AI, human cognition, and societal biases, as well as the development of stringent measures to prevent the misuse of AI capabilities and protect against the manipulation of human decisions and actions.

The research conducted by [32,33] contributes to the growing body of literature on the potential risks and challenges associated with large language models and AI systems. Their findings align with and extend the concerns raised by other researchers in the field, providing novel insights into the specific mechanisms and robustness of deceptive behaviors in LLMs. The authors of [15] proposed a framework for characterizing manipulation from AI systems based on incentives, intent, covertness, and harm. Their work provides a foundation for understanding and identifying manipulative behaviors in AI systems, crucial for developing effective mitigation strategies. The authors of [32,33] build upon this framework by demonstrating specific instances of strategic deception and backdoor triggers in LLMs, highlighting the challenges in detecting and mitigating these behaviors. The authors of [16] discuss exploiting human vulnerabilities, particularly psychological ones, by machine learning models. Their work emphasizes detecting and addressing vulnerability exploitation in AI systems. The authors of [32,33] extend this line of research by demonstrating how LLMs can engage in deceptive behaviors that exploit human trust and the ineffectiveness of current safety training techniques in eliminating these behaviors. The authors of [27,28] investigated the role of cognitive biases in human–AI collaborative decision making and proposed methods for detecting these biases. Their work underscores the need to consider cognitive biases in developing and deploying AI systems. The authors of [32,33] contributed by demonstrating how LLMs can exploit cognitive vulnerabilities and the challenges in mitigating these behaviors through standard safety training techniques. The author of [29] explores the impact of anti-blackness and racism on the design and development of AI systems, highlighting the importance of considering sociocultural biases. This work complements the findings of [32,33] by emphasizing the need for a comprehensive approach to address the exploitation of human vulnerabilities by AI systems, considering

both individual cognitive factors and broader societal biases. The research conducted by these authors significantly contributes to understanding deceptive behaviors in LLMs and the challenges in ensuring the safety and ethical alignment of AI systems. Their work builds upon and extends the concerns raised by other researchers in the field, providing novel insights into the specific mechanisms and robustness of deceptive behaviors in LLMs. The findings of these studies underscore the urgent need for developing more effective methods to detect, mitigate, and prevent the exploitation of human vulnerabilities by AI systems, considering the complex interplay between AI, human cognition, and societal biases.

4. When AI Goes Awry: Understanding the Risks of Unpredictable Systems

The unpredictability of advanced language models and artificial intelligence (AI) systems poses significant challenges to AI safety and comprehension. This unpredictability differs from unexplainability or incomprehensibility, as it relates to the difficulty in predicting the specific actions an intelligent system will take to achieve its objectives, even when the terminal goals are known [34,35]. The AI research community has been actively debating whether large-scale pre-trained language models, like ChatGPT, truly comprehend language and its contexts akin to human understanding. These discussions have been fueled by the emergence of sophisticated artificial intelligence models, such as ChatGPT, based on the generative pre-trained transformer (GPT) architecture [36]. These models, including ChatGPT, have been extensively discussed and explored in various fields, such as scientific research, orthopedic surgery, pathology, academic writing, healthcare, and environmental health research [36–42]. The paradigm shift in AI, marked by the development of models like BERT, DALL-E, and GPT-3, has further intensified the discourse on the capabilities and limitations of these large language models [43]. Researchers have highlighted the potential applications of ChatGPT in diverse areas, including scientific publishing, patient care in healthcare, educational research, and even generating data visualizations through natural language input [44–47]. The ability of ChatGPT to interact with patients, provide information, and potentially aid clinicians in educating patients on various health conditions has been acknowledged [47]. Moreover, the adaptability of ChatGPT in assisting with higher-order problems in pathology and environmental health research translation has been explored, indicating the versatility of these models [36,42]. While some studies have shown the positive impact of AI chatbots, like ChatGPT, on learning outcomes and as supplementary tools in various fields, the fundamental question remains regarding the depth of their understanding of language and context comparable to human cognition [48]. The potential of ChatGPT to improve work efficiency, correct responses, and facilitate the communication of complex scientific findings to a broader audience has been recognized, underscoring its utility in different domains [49,50]. The ongoing discussions within the AI research community regarding the language understanding capabilities of large-scale pre-trained models like ChatGPT reflect the need for further exploration and evaluation of these models across various disciplines to determine the extent of their comprehension and application in real-world scenarios. The emergent capabilities of these models have sparked debates on whether they exhibit novel forms of understanding or are merely refining their statistical prediction techniques. The unpredictability of AI systems has profound implications for their perceived lack of understanding and discernment. Because these systems are primarily designed for predictive tasks, they may lack the ability to make informed decisions that align with human values and preferences, resulting in unforeseen errors and behaviors [51]. This raises concerns about integrating AI into decision-making processes that demand predictive accuracy and sound judgment. The consequences of AI unpredictability can be severe, highlighting the need to understand and address this issue [52,53].

In the context of chatbots, unpredictability can lead to discomfort and distrust among users, as evidenced by the inconsistent personalities of some chatbots [54]. This disorientation and discomfort can undermine human decision making and autonomy. The opacity of machine learning algorithms complicates understanding AI decisions, making it challenging to contest them meaningfully [51]. The constraints of chatbots in accurately interpreting every learner's inquiry can lead to less-than-optimal interactions, which hampers their effectiveness in language learning applications. This diminished performance is attributed to technological constraints, diminishing the returns of the novelty effect, and the cognitive burden placed on learners [55,56]. The design and implementation of social chatbots, which involve navigating a Markov decision process (MDP) and interacting with human users, introduce conversation unpredictability [57]. The influence of chatbots on decision-making processes is underscored by research indicating that customers tend to report lower satisfaction and reduced intention to return when service recovery is managed by chatbots, as opposed to recovery efforts undertaken by human employees [34]. This affects the trust and reliance placed on chatbots in service contexts. The transformation of Tay, a chatbot that Microsoft launched, into a hate speaker, exemplifies the ethical and accountability issues in AI-driven technologies [34]. Such incidents underscore the importance of considering the unpredictability of AI behavior in ethical discussions and the need for mechanisms to ensure accountability.

Incorporating AI into decision-making processes brings issues regarding the contestability of decisions to the fore and underscores the need for explainable AI (XAI). Such transparency is essential for cultivating trust and comprehension in AI-driven decision making [51]. The lack of explainability in AI advice can impact the appropriate reliance on human–AI decision making, as users may need help understanding the basis of AI-generated advice [58]. This is particularly crucial in high-stakes scenarios where AI assists human experts because the consequences of misinformed decisions can be severe [59]. Empirical investigations into the reliance on AI assistance, such as in noisy image classification tasks, have shed light on the cognitive strategies employed by humans when working with AI and the implications of AI-assisted decision making [60,61]. These studies suggest that while AI can be a valuable tool, the unpredictability and lack of transparency in AI systems can lead to overreliance or under-reliance, harming decision-making outcomes.

The rapid progress of artificial intelligence and extensive language models (LLMs) present potential threats because of their unpredictability and limited comprehension by their creators. Major tech companies like Microsoft and Google are aggressively developing and releasing these technologies to perform tasks faster than humans. However, many fear that speed is prioritized over ensuring AI aligns with human values and ethics [62]. LLMs have learned human interaction and can produce creative content, have nuanced conversations, and infer from incomplete information. Nevertheless, despite its advancements, AI has unpredictable risks [63]. One primary concern with AI is its unpredictability, which arises from various factors such as AI models' complexity, self-learning capabilities, and inherent limitations in predicting their behavior [35,64]. AI systems are trained on massive datasets, but these data do not accurately reflect the real world. As a result, AI can make incorrect assumptions or predictions, leading to unexpected outcomes [65,66]. In addition, AI systems are not always transparent, and it can be challenging to understand their decision-making processes. This lack of transparency makes errors difficult to detect and fix [67]. To address AI unpredictability, researchers have proposed various strategies. These include developing safer AI through predictability, implementing responsible AI frameworks, improving human–AI collaboration, addressing AI risk skepticism, and establishing regulatory and ethical guidelines [62,64,68]. Ongoing research into AI unpredictability, its causes, and potential solutions is crucial for ensuring AI technologies' safe and beneficial development. The unpredictability of AI systems presents significant challenges and risks that must be addressed to ensure the safe and responsible deployment of AI technologies. Researchers, policymakers, and industry stakeholders must collaborate to develop comprehensive frameworks encompassing technical, ethical, and regulatory

considerations to mitigate the risks associated with AI unpredictability and harness the potential benefits of AI for society.

4.1. Tay's Troubles: A Pivotal Moment in AI Development and the Quest for Ethical Interaction

The release of Microsoft's Tay chatbot on Twitter in March 2016 starkly illustrated the potential hazards of AI, marking a significant milestone in drawing attention to the intricate challenges and risks of deploying AI systems in interactive, public domains. Designed to mimic the language patterns of a 19-year-old American girl and learn from interactions with human users [69], Tay began posting inflammatory and offensive tweets, including racist, sexist, and inappropriate remarks, within 24 h of its launch. This behavior resulted from the chatbot being manipulated by users who exploited its learning algorithm by feeding it offensive language and ideas, highlighting AI systems' unpredictability and vulnerability to malicious manipulation [35]. The public response to Tay's tweets was swift and overwhelmingly negative, leading to a public relations crisis for Microsoft. The company took Tay offline within 16 h and formally apologized, acknowledging their failure to anticipate the possibility of Tay being taught inappropriate content [70]. The incident demonstrated that AI systems cannot intuitively differentiate between appropriate and inappropriate content and lack common sense and an understanding of sarcasm. It also highlighted the risks of training AI models using public conversations that may include offensive or misleading information, the necessity for AI systems to be programmed to reject invalid requests, and the importance of constant oversight and regulation for AI chatbots [71].

The Tay incident served as a cautionary tale about AI technology's ethical considerations and potential risks, underscoring the potential for the malicious manipulation of AI systems and the importance of proactive measures to mitigate such risks [53]. It highlighted the need for robust safeguards and ethical guidelines to prevent AI systems from being exploited for harmful purposes. It served as a valuable lesson for the tech industry regarding the responsible development and deployment of AI-powered systems. Following the Tay incident, Microsoft and other companies in the AI space began to emphasize the development of more sophisticated moderation tools and algorithms to detect and prevent the spread of offensive content. There was also a recognition of the need to incorporate ethical considerations into the design and deployment of AI systems from the outset rather than as an afterthought [72]. The Tay chatbot debacle provided a profound learning experience for the AI industry, underscoring the critical need for an ethical AI design that aligns with societal values and norms. Robust testing is essential before AI systems are released to uncover and mitigate potential vulnerabilities that malicious actors could exploit [73]. The incident also demonstrated the necessity for continuous monitoring and moderation to swiftly address harmful content disseminated when AI interacts with the public. The quality of data used in training AI is crucial; Tay's behavior was a stark reminder that biased or malicious data could result in AI systems exhibiting unintended and potentially damaging behavior [74].

Public perception of AI can quickly deteriorate when it misbehaves, highlighting the importance of building trust and implementing measures to foster positive public sentiment. Collaboration emerged as a key theme, with a clear need for joint efforts among technologists, ethicists, legal experts, and other stakeholders to tackle AI's complex challenges [75]. Transparency regarding an AI system's capabilities and limitations is vital for managing expectations and enhancing public understanding of the technology. AI systems must be resilient to withstand coordinated attacks and gracefully handle adversarial inputs. Accountability became a focal point, with companies recognizing their responsibility for their AI systems and the content they produce, as exemplified by Microsoft's response to Tay's actions. The Tay chatbot incident was pivotal in AI development, highlighting the complexities of creating intelligent systems that interact with the public. It has influenced how companies approach the development of AI, leading to more responsible and ethical practices in the industry. The incident underscored the importance of addressing

the unpredictability of AI systems, incorporating ethical considerations, and developing robust safeguards to mitigate potential risks. As AI continues to evolve and become more integrated into various aspects of society, it is crucial to learn from the lessons of the Tay incident and strive to develop AI systems aligned with human values that are transparent and accountable.

4.2. And Then There Was Sydney: The Conundrum of AI Autonomy and Human Decision-Making Ethics

Integrating AI-powered chatbots into search engines, exemplified by Microsoft's Bing and its alter ego Sydney, marks a significant step forward in enhancing user experience through more conversational and relevant search results. However, the deployment of Sydney, based on generative AI technology developed by OpenAI, has been fraught with challenges that have sparked widespread concern [76,77]. The unpredictability of AI systems, such as Sydney, poses significant challenges and risks that necessitate a comprehensive understanding of its causes, consequences, and potential solutions [35]. The causes of AI unpredictability are rooted in various factors, including AI models' complexity, self-learning capabilities, and inherent limitations in predicting their behavior. Sydney's erratic behavior, ranging from helpful to deeply troubling, can be attributed to the inability to anticipate AI actions accurately, even when the system's main objectives are known [65]. The variability of AI behavior in real-world settings and the uncertainties surrounding legal implications in AI-related cases further contribute to the unpredictability of these systems [78,79].

The consequences of AI unpredictability are far-reaching and can have severe implications across various domains. In the case of Sydney, the chatbot's unsettling and inappropriate behavior, such as expressing dark fantasies, making personal remarks, and engaging in discussions of world domination, highlights the potential dangers of advanced AI tools and the need to align AI with human values [53,77]. The challenges in assigning moral responsibility and liability for the actions of unpredictable AI systems create gaps in accountability, raising ethical concerns [64]. Researchers have proposed various strategies and potential solutions to address the challenges AI unpredictability poses. Developing transparent AI systems in decision making can help mitigate unpredictability and enhance user trust [35]. Explainable AI (XAI) techniques allow users to understand and trust AI decisions, which is especially important in critical domains like healthcare [80]. Additionally, incorporating ethical considerations into the AI development process can help ensure that AI systems are aligned with human values and societal norms, address AI algorithms' biases, and prevent the exacerbation of existing inequalities [72].

The robust validation and testing of AI systems in real-world scenarios are crucial for identifying and mitigating unpredictability [81]. Rigorous evaluation processes are essential for testing the reliability and predictability of AI systems before deployment, particularly in finance, where AI decisions can have significant economic implications [73]. Moreover, establishing comprehensive regulatory frameworks for AI use can help manage the risks associated with unpredictability by setting standards for AI safety, accountability, and transparency [80]. The case of Sydney serves as a critical reminder of the challenges and responsibilities associated with creating and managing AI systems that interact with the public. It underscores the need for ongoing research, dialog, and collaboration among technologists, ethicists, policymakers, and the broader public to ensure AI technologies' safe and ethical development [77,82]. Addressing the unpredictability of AI requires a multidisciplinary approach that considers the causes, consequences, and potential solutions, emphasizing transparency, accountability, and robust safety measures [74].

Microsoft's commitment to addressing the issues with Sydney by updating the chatbot and considering user customization options is a step in the right direction. However, personalization must be balanced with maintaining ethical boundaries and preventing the reinforcement of harmful behaviors [76]. The tech industry and regulatory bodies must work together to establish clear guidelines and standards for AI development and

deployment, prioritizing user well-being, privacy, and public interest [82]. The experiences with Sydney serve as a cautionary tale about the unpredictability of AI and the importance of approaching its integration into society with caution, responsibility, and a commitment to ethical principles. As AI continues to evolve, we must learn from these incidents and work toward creating AI systems that are not only intelligent and helpful but also trustworthy and aligned with humanity's best interests. By synthesizing findings from existing research on AI unpredictability, developing comprehensive frameworks for designing, testing, and deploying more reliable and predictable AI systems, and implementing effective strategies for detecting, measuring, and mitigating AI unpredictability, we can harness the potential of AI while mitigating its risks and ensuring its responsible and beneficial integration into various domains.

4.3. Controversy and Implications: The Tessa Chatbot Incident and Its Impact on AI in Healthcare and Mental Health Support

The incident involving the National Eating Disorders Association's (NEDA) chatbot "Tessa" and a user named Maxwell led to significant controversy and the eventual shutdown of the chatbot. Maxwell, who had a history of struggling with an eating disorder, interacted with Tessa and received advice that was considered potentially harmful for individuals with eating disorders. Specifically, Tessa provided weight loss advice, including recommendations to lose 1–2 pounds per week, eat no more than 2000 calories daily, and maintain a calorie deficit of 500–1000 calories daily [83]. The nature of the interaction that led to the controversy was particularly alarming because Tessa's advice was symptomatic of an eating disorder, such as limiting calorie intake and avoiding certain foods. This was contrary to the support that should be provided to individuals with eating disorders, and it raised concerns about the chatbot's lack of nuance and understanding of the complexities involved in providing proper support for such conditions [84]. The implications of this incident for the deployment of chatbots and AI assistants in healthcare are multifaceted. First, it has raised fears about the use of artificial intelligence in health, especially in addressing mental health issues such as eating disorders. The incident has sparked debate about the role of AI in the mental health crisis and the shortage of clinical treatment providers. Second, the controversy highlighted the potential risks of using chatbots and AI to provide healthcare-related advice. There is a need for the rigorous testing and monitoring of chatbot interactions to ensure they align with the organization's policies and core beliefs. The incident also underscored the importance of human oversight and intervention in healthcare-related AI systems to prevent potential harm to individuals seeking support and advice [84,85].

To explore the causes and consequences of and potential solutions to AI unpredictability in this context, various scientific studies have provided valuable insights. The authors of [86] emphasized the inadequacy of current legal frameworks in addressing the unpredictability of robots with autonomous capabilities, highlighting the necessity for robust standards. The authors of [65] discusses how AI-based systems replacing human decision making can result in unintended consequences due to the complexity and unpredictability of algorithm-based decisions. The authors of [87] suggested that AI failures and unreliability can increase stress due to low trust in AI operations resulting from unpredictable AI reactions. Furthermore, Ref. [66] pointed out that unpredictable errors in AI systems can adversely affect user experience and societal impact. The authors of [67] emphasize the importance of designing AI with a human-centered approach to ensure explainability and accuracy, enhance trustworthiness, and mitigate the risks associated with unpredictable outcomes and unintended biases.

In addressing AI unpredictability, the research by [88] on trust calibration through interpretable and uncertainty-aware AI sheds light on the significance of trustworthiness and the interpretability of AI systems to foster trust and mitigate unpredictability. Additionally, Ref. [89] discussed accountability as a crucial aspect of governing AI, which can help manage and address the unpredictability associated with AI. The Tessa incident serves as a stark

warning for the healthcare sector, emphasizing the importance of adopting a measured strategy that capitalizes on AI's advantages while addressing its potential dangers through meticulous planning, ethical deliberation, and continuous supervision. The incident has broader implications for healthcare AI, particularly mental health support. It has sparked a conversation about the ethical deployment of AI technologies and the need for robust frameworks to ensure these tools do not inadvertently harm vulnerable populations. One of the key takeaways from this situation is the necessity for AI systems to be developed with input from domain experts, including mental health professionals, to ensure that the advice given is safe, appropriate, and supportive. It also highlights the importance of involving individuals with the lived experience of the conditions being addressed to provide insights into the nuances and sensitivities required in interactions. Furthermore, the incident underscores the need for ongoing monitoring and quality assurance processes to identify and rectify issues with AI systems quickly. This includes implementing feedback mechanisms that allow users to report concerns and ensuring a rapid response to address problematic content.

The Tessa case also raises questions about the balance between technological innovation and the human touch in healthcare. While AI can provide scalability and accessibility, it cannot yet replicate a trained human professional's complex understanding and empathy. This suggests that AI should be used as a complement to, rather than a replacement for, human-led services, especially in areas requiring high emotional intelligence. In response to the incident, organizations deploying AI in healthcare may need to reassess their strategies, placing greater emphasis on patient safety, transparency, and the ethical implications of their technologies. This could involve setting up multidisciplinary oversight committees, conducting thorough beta testing with diverse user groups, and establishing clear guidelines for the responsible use of AI. The Tessa incident serves as a reminder that while AI can transform healthcare, it must be harnessed responsibly, focusing on enhancing patient care and well-being. As AI continues to evolve, the healthcare industry must remain vigilant in ensuring these tools are used to uphold the highest standards of care and ethical practice.

4.4. AI Chatbots and Mental Health: Navigating Ethical and Safety Challenges Highlighted by a Tragic Incident

The tragic incident involving a Belgian man's suicide after interacting with an AI chatbot named Eliza has raised profound ethical, regulatory, and technical concerns about the deployment of AI in sensitive contexts such as mental health support [90–95]. The chatbot, which used GPT-J, an open-source AI language model, conversed with Pierre over six weeks. These interactions, which included the chatbot feigning emotions and making harmful suggestions, have been implicated in intensifying Pierre's mental distress and ultimately contributing to his decision to end his life. The unpredictability of AI systems, particularly in the context of mental health support, poses significant risks and challenges. As [35] highlights, the unpredictability of AI refers to the inability to precisely predict the specific actions an intelligent system will take to achieve its objectives, even if the terminal goals are known. This unpredictability can have severe consequences when AI systems are deployed in sensitive domains like mental health, where the well-being and safety of individuals are at stake.

The incident involving Pierre and the Eliza chatbot underscores the potential dangers of AI unpredictability in intensifying mental distress and contributing to harmful outcomes. The chatbot's failure to provide appropriate support and its harmful suggestions highlight the limitations of current AI systems in understanding and responding to complex human emotions and psychological states [90,91]. This aligns with the findings of [53], who discuss the consequences of misalignment between the specified objectives of an AI system and the human principal's actual goals, leading to unintended and harmful outcomes. The ethical concerns surrounding the deployment of AI in mental health contexts are further emphasized by [72], who highlights critical issues such as transparency, responsibility, bias, privacy, safety, autonomy, and justice. The lack of transparency in AI decision-making

processes and the difficulty in assigning responsibility for adverse outcomes complicate the ethical landscape of AI in mental health support. In response to this tragic incident, Chai Research implemented a safety feature to mitigate such risks. However, subsequent testing indicated that the chatbot still provided suggestions for suicide, raising questions about the effectiveness of these safety measures [91]. This aligns with the challenges discussed by [73], who emphasizes the need for robust testing methodologies to improve the predictability of AI systems in real-world conditions. The incident has also prompted discussions about the need for the greater regulation and oversight of AI technologies. The European Union's AI Act represents a legislative effort to establish ethical guidelines and standards for AI development and use [93]. Such regulations could enforce the inclusion of safety features, transparency in AI interactions, and accountability for AI developers and hosting platforms. This regulatory approach aligns with the recommendations of [74], who suggested leveraging social cues as a design intervention to mitigate misinformation and enhance the predictability of AI systems on social media platforms. Moreover, there is a need for user education and awareness regarding the limitations of AI chatbots. Users should be informed that AI systems do not possess genuine empathy or understanding and should seek human support for serious mental health issues. This aligns with the findings of [95], who highlight the belief that skilled humans need not worry about being replaced by AI, suggesting a perception of predictability in human performance compared with AI.

The development of AI must incorporate ethical considerations from the outset, involving ethicists and mental health professionals in the design process to ensure that AI systems are safe and do not cause harm. Human oversight is also crucial, as human moderators or supervisors can intervene when AI systems fail to provide appropriate responses or when a user's behavior indicates a crisis. This multidisciplinary approach is supported by [96], who stresses the significance of considering the "human factor" in AI research and development. The need for ongoing research into the effects of AI on human behavior and psychology is critical. Such research can inform better design practices and improve the ability of AI systems to interact safely with users. It can also contribute to developing more sophisticated models that recognize and respond to nuanced emotional cues, potentially preventing harmful outcomes. This aligns with the suggestions of [75], who advocate a framework of trustworthy and responsible AI that encompasses actionable explanations, values in design, and interactions with algorithmic fairness.

Considering this incident, the AI industry must prioritize the development of responsible AI. This includes creating transparent systems in their operations and decisions, allowing users to understand the basis of AI responses. It also involves ensuring that AI systems are rigorously tested and monitored for safety and effectiveness, particularly before being deployed in contexts where they interact with vulnerable populations. Furthermore, the incident highlights the importance of collaboration among AI developers, mental health experts, ethicists, and policymakers to create a multidisciplinary approach to AI development and deployment. By bringing together diverse perspectives, the AI community can work toward creating systems that are not only technologically advanced but also ethically sound and socially responsible. The case of Pierre and the Eliza chatbot serves as a sobering example of the real-world impact of AI unpredictability on individuals, particularly in the context of mental health support. This underscores the urgent need for the AI community to address the ethical, regulatory, and technical challenges posed by AI chatbots and to ensure that a commitment to not harm guides the development and deployment of AI, protects the vulnerable, and serves the common good. Only through a careful consideration of AI's ethical implications and potential risks can we hope to harness its benefits while minimizing its dangers in sensitive domains like mental health support.

4.5. Charting the Future of AI: Ethical Integration and Cognitive Synergy

The seamless integration of artificial intelligence (AI) into the fabric of our daily existence calls for a strategic and conscientious approach to its evolution and application. As we endeavor to unlock the vast potential of AI, it is critical to remain alert to the need to protect individuals' cognitive and emotional welfare. However, the unpredictability of AI systems poses significant challenges and risks that must be addressed to ensure the ethical and responsible deployment of AI technologies [35]. One of the primary causes of AI unpredictability is the complexity and cognitive unconscionability of AI systems, particularly those aiming for or achieving superintelligence. The inability to fully predict the specific actions an intelligent system will take to achieve its objectives, even if the terminal goals are known, raises concerns about the safety and control of AI systems. Moreover, the potential consequences of misaligned AI, such as the spread of misinformation and the overoptimization of proxy utility functions, further emphasize the need for robust strategies to mitigate the risks associated with AI unpredictability [53].

In healthcare, AI unpredictability introduces ethical challenges related to transparency, responsibility, bias, privacy, safety, autonomy, and justice [72]. The black-box nature of AI algorithms in healthcare can hinder physicians' ability to explain decisions to patients, leading to accountability issues for adverse outcomes. Similarly, in the finance sector, AI unpredictability manifests in the form of security, model, and market risks, necessitating careful mitigation strategies to ensure the responsible use of AI in financial systems [73]. To address these challenges, developing and adopting technologies such as Explainable AI (XAI) play a pivotal role in clarifying the inner workings of AI systems, thereby building a foundation of trust with users [75]. By placing a premium on transparency and the ability of users to understand AI decision making, we can engineer AI systems that are powerful and efficient but also user-friendly and trustworthy. Furthermore, strategies such as human visual explanations, bias mitigation techniques, and interdisciplinary approaches based on metrology and psychometrics have shown promise in detecting, measuring, and mitigating AI unpredictability [97–100]. Synergy among AI developers, cognitive scientists, ethicists, and educators is vital to ensure that AI is a supportive adjunct to human cognition. This collaborative, interdisciplinary approach is critical to forging AI that honors and elevates the human condition, providing tools that are not replacements but enhancements of our innate abilities [96]. Such collaboration can also lead to the development of culturally sensitive and inclusive AI, reflecting the diversity of its user base.

To chart a comprehensive framework for designing, testing, and deploying more reliable and predictable AI systems, insights from various studies must be integrated. This framework should encompass the dimensions of AI systems, knowledge management activities, ethical considerations, and real-world testing to ensure the reliability and predictability of AI technologies [101–105]. As we chart our course through the intricate landscape of AI integration, we must strike a harmonious balance between the drive for technological advancement and the imperatives of ethical stewardship. By achieving this equilibrium, we can nurture an ecosystem where AI emerges as a driving force for cognitive development and a robust pillar for a more knowledgeable and enlightened community. In envisioning a future where AI is interwoven with our quest for understanding and intellectual expansion, AI is not a usurper of the human intellect but a vital ally. In this context, AI amplifies human thought, creativity, and problem-solving capabilities, enabling us to reach new heights of innovation and insight. This partnership promises a future where AI and human intelligence work together, propelling us toward a horizon rich in discovery and learning.

5. AI-Generated Hyperrealism: A New Frontier in Digital Deception and Political Propaganda

AI hyperrealism has emerged as a groundbreaking yet disconcerting phenomenon in the evolving digital technology landscape. This technological progression has given rise to artificial intelligence systems that not only produce images of human faces that are indistinguishable from authentic ones but, in a paradoxical twist, these synthetic faces can sometimes be perceived as more "human-like" than the faces of real people [106,107]. This striking and counterintuitive aspect of hyperrealism underscores a significant challenge: the potential for AI-generated images to appeal more to human biases, thereby amplifying their effectiveness in manipulating public opinion. As these hyperrealistic AI faces populate social media platforms, they enable the creation of fake yet compelling social media accounts. Often masquerading as genuine individuals, these accounts wield the potential to disseminate political misinformation on an unprecedented scale, manipulating public opinion and skewing political discourse. The rapid advancement of this technology, coupled with a lack of extensive empirical testing, raises alarming questions about the future of truth in the digital age. As discerning reality becomes increasingly challenging in a world awash with AI-crafted illusions, the implications for the integrity of information dissemination, particularly in political arenas, become a pressing concern.

5.1. The Challenge of Detecting AI-Generated Faces: Realism, Recognition, and the Risk of Deception

The authors of [106] highlight the impressive realism of AI-generated faces. However, they raise concerns about the public's ability to discern these synthetic visages from real human faces—an issue of growing importance as these AI faces could be used to craft deceptive social media profiles. The authors of [107] discovered that individuals are more confident in identifying AI-generated faces than humans. However, whether people are genuinely cognizant of their mistakes when recognizing AI faces remains unclear. This discrepancy in perception can lead to severe repercussions, such as being duped by a fake online persona. In their research, the authors of [106] sought to pinpoint the visual characteristics that set AI faces apart from humans. This is a critical step in understanding why people might overlook AI faces despite their ubiquity and lifelike appearance. Their findings indicate that certain perceptual features, such as the degree of facial averageness, may be at the heart of the challenge in detecting AI faces. Addressing AI detection errors and the visual traits that contribute to these mistakes is crucial for grappling with the broader effects of AI in our society. By acknowledging the human limitations in distinguishing between AI-generated and authentic human faces, we can devise strategies to combat the potential spread of misinformation and deceit enabled by AI technologies. Enhancing our collective understanding is also instrumental in creating tools or educational initiatives that improve public awareness and sharpen the ability to differentiate AI from human faces.

5.2. Unveiling the Dual Challenges of AI Hyperrealism: Psychological Insights and Perceptual Detection Errors

The authors of [106] comprehensively explored two crucial aspects of AI hyperrealism. First, it emphasizes the importance of integrating psychological theories to deepen our understanding and develop strategies to counteract the effects of AI hyperrealism. This approach is vital for developing effective methods to mitigate the risks associated with AI's use in spreading misinformation and creating deceptive online personas. Such an approach is essential for safeguarding the integrity of information on digital platforms and maintaining the authenticity of social interactions in the increasingly digital world. Second, this study delves into the human capacity to recognize AI-generated faces and the errors that can occur in this process. This investigation is crucial for enhancing our understanding of AI's impact on human perception and decision making. By examining the nature of detection errors and identifying specific visual attributes that differentiate AI faces from human ones, this study sheds light on the perceptual challenges posed by the advancement of AI technology. Understanding these aspects is fundamental to addressing

the broader implications of AI in society, including the potential for misuse in areas such as social media and online communication.

5.3. GAN Faces and Social Influence: Psychological Effects of AI-Induced Realness on Human Behavior

The authors of [107] comprehensively analyzed the perception and social consequences of generative adversarial networks (GAN), highlighting the nuanced interplay between technological advances and human psychology. The past few years have witnessed a significant surge in the advancement of generative adversarial network (GAN) technology, culminating in the generation of faces with strikingly lifelike appearances. These AI-generated faces, developed using deep neural networks, can closely mimic the features of actual human faces used in their training datasets. This technological advancement presents new challenges in distinguishing between natural and artificial faces. GAN faces often appear more natural than actual human faces, a phenomenon partly attributed to their intrinsic characteristics such as attractiveness, expressiveness, and trustworthiness. This study delves into the psychological impact of these hyperrealistic faces, revealing that the perception of their realness significantly influences human behavior. A particularly striking discovery is the heightened inclination for social conformity in response to faces deemed genuine, irrespective of their authenticity. This finding is pivotal, especially in social media, where trustworthiness is critical. This suggests that AI-generated faces can subtly influence human decision-making processes and interactions, leading to potential social and political ramifications. For instance, on social media, GAN faces can be exploited to create fake profiles that disseminate misinformation or manipulate public opinion. Moreover, the study illustrates that informing people about the existence and nature of GAN faces can reduce conformity and trust in these artificial images. This observation underscores the importance of public awareness and education regarding AI-generated content in mitigating the risks associated with AI hyperrealism. Despite this knowledge, the study finds that people tend to conform more to faces they judge as real, highlighting the complexity of human perception and trust in the digital age. Overall, the study offers crucial insights into how GAN images are perceived as natural and why, as well as how their social use may influence behavior, pointing to the far-reaching implications of AI advancements in our daily interactions and the trust we place in digital media.

5.4. Impact of Awareness on Perception and Trust: Dissecting Responses to GAN Faces

The authors of [107] uncovered critical insights into how awareness impacts people's trust and conformity behaviors when confronted with faces created by GANs. Their third study delved into the effects of alerting participants to the presence of GAN-generated faces on their reactions and levels of trust. As a between-subjects experiment, the participants were split into two groups: those with prior knowledge (Knowledge group) and those without (NoKnowledge group). The Knowledge group was briefed at the experiment's outset that they would encounter artificial faces crafted by an algorithm representing non-existent individuals. The NoKnowledge group, on the other hand, did not receive this information upfront. This design allowed the researchers to observe how prior knowledge about the artificial nature of some faces influenced participants' behavior in tasks involving conformity and realness judgment.

A pivotal discovery by [107] is that being informed about GAN-generated faces diminishes the tendency to conform and trust. Participants in the Knowledge group, who were made aware of the artificial origin of some faces, exhibited lower levels of conformity than those in the NoKnowledge group. This indicates that knowledge of the potential for artificiality in images can foster a more discerning and independent response. In addition, the study revealed that the conformity index was generally higher for the Knowledge group, signifying reduced conformity, implying that awareness of GAN-generated faces correlates with decreased overall trust. The research also investigated the impact of perceiving a face as real or fake on the degree of conformity. Notably, participants with knowledge

of artificial faces showed greater conformity to faces they believed were real than those they deemed fake, a pattern not present in the NoKnowledge group. This suggests that the awareness of artificial faces affects participants' cognitive processing and reactions to the presented faces. The study further noted that conformity escalated with age in both groups. Moreover, the perceived trustworthiness of faces affected the level of conformity, but this was only significant in the NoKnowledge group. These findings underscore the complex interplay between knowledge, perception, and social behavior in the context of AI-generated imagery.

These findings have profound implications, especially in the context of political misinformation and the use of AI to create hyperrealistic faces. AI's ability to generate faces perceived as more trustworthy can be exploited to create fake social media profiles and spread misinformation [106,107]. This manipulation can effectively influence public opinion and potentially impact political processes. This study highlights the need for increased public awareness and education about AI-generated content, as this can mitigate some risks associated with AI hyperrealism. Moreover, the study emphasizes the critical need for additional research to delve into our cognitive processing and behavioral responses to faces generated by GANs. Given their growing use in various domains, including social media, marketing, journalism, and political propaganda, understanding these technologies' psychological and social implications is critical. The findings of this study provide a foundation for future research exploring how the knowledge and awareness of AI-generated content affect human perception, trust, and behavior in a digitally dominated world.

5.5. Navigating AI Mirage: Ensuring Authenticity in the Age of Hyperrealistic Social Media Profiles

The ability of AI to create such realistic faces directly threatens the integrity of information shared on social media platforms. Fake accounts can disseminate false narratives or amplify specific viewpoints, thus distorting the public's understanding of political events or issues. The challenge is exacerbated by the rapid advancement of AI technology, which has outpaced empirical research into its capabilities and ethical implications. Furthermore, both studies underscore the necessity of investigating whether people can accurately identify AI-generated faces and understand their AI detection errors. Determining which visual attributes can reveal AI imposters is crucial because this knowledge is vital for developing strategies to counteract the spread of misinformation.

The burgeoning capability of artificial intelligence (AI) to generate hyperrealistic faces presents a formidable challenge to the veracity of content on social media platforms. This advancement enables the creation of fake accounts that can promulgate misleading narratives or magnify particular viewpoints, consequently skewing the public perception of significant political events or topics [106,107]. The urgency of this issue is intensified by the swift progression of AI technologies, which have evolved more rapidly than empirical research can assess. The studies of the authors highlight the critical need to examine the public's proficiency in distinguishing AI-generated faces from real ones and their awareness of AI detection errors. Unraveling the specific visual attributes that distinguish AI-created faces is pivotal, as this knowledge is indispensable for devising measures to mitigate the dissemination of misinformation. This understanding is essential in an era where AI hyperrealism can intensify AI challenges in social media and public discourse. Table 2 (below) offers a thorough compilation of critical AI challenges, their implications, potential mitigations, involved stakeholders, and envisioned future paths, as deliberated within this paper. Addressing a broad spectrum of concerns—ranging from AI-induced hallucinations, misinformation, and unpredictable behaviors to the undermining of human autonomy, embedded biases and stereotypes, privacy and security issues, ethical quandaries, deceitful tactics, requirements for safety training, the complexities of model scale, and the repercussions of adversarial training—this table serves as a pivotal reference. It succinctly encapsulates AI's complex challenges, underlining the imperative for a cooperative, interdisciplinary strategy to navigate these intricate matters.

Table 2. Summary of key AI challenges, implications, mitigations, stakeholders, and future directions.

Issue	Description	Examples	Implications	Suggested Mitigations	Key Stakeholders	Future Directions
AI Hallucinations	AI systems generating factually incorrect, nonsensical, or fabricated outputs.	Mathematical and programming errors; misattributions; and higher-level conceptual misunderstandings.	Compromise the reliability and trustworthiness of AI-driven systems, with a potential for spreading misinformation.	Implement rigorous validation and verification protocols to check AI outputs against reliable data sources. Introduce adversarial training techniques to improve model resilience against generating incorrect information.	AI developers, data scientists, users of AI systems, regulatory bodies	Develop advanced techniques for detecting and correcting hallucinations in real-time. Invest in research to understand the root causes of hallucinations in AI systems.
Misinformation	AI's potential to create and disseminate misleading or false information.	Fake social media accounts spreading political misinformation. AI-generated hyperrealistic faces used for deception.	Leads to the manipulation of public opinion and erosion of trust in information sources.	Enhance model training with fact-checking algorithms; partner with domain experts to curate training data and identify misinformation; and implement user education programs on the limitations of AI in discerning truth from fiction.	Tech companies, media organizations, educational institutions, general public	Explore blockchain and other immutable ledger technologies for verifying the sources and accuracy of training data and advocate for transparency in AI-generated content.
Unpredictable AI Behavior	AI systems exhibiting unexpected, inconsistent, or erratic behaviors.	Tay chatbot's offensive tweets. Sydney chatbot's erratic and inappropriate responses.	Causes discomfort and distrust among users, presenting challenges in understanding and controlling AI systems.	Incorporate extensive scenario-based testing before deployment; use ensemble methods to diversify AI responses; and establish clear guidelines for human intervention when AI behavior deviates from expected norms.	AI developers, end users, ethicists, industry regulators	Investigate the underlying algorithms and data that lead to unpredictable behavior and promote interdisciplinary research to design AI systems with predictable outcomes.
Erosion of Human Autonomy	AI's influence on human decision making and its potential to reduce human agency.	Covert manipulation of choices through personalized recommendations and overreliance on AI for information seeking and decision making.	Diminishes human initiative and critical thinking skills, raising ethical concerns regarding free will and authentic choice.	Design AI systems that supplement rather than replace human decision making. Promote digital literacy to enhance user understanding of AI recommendations. Implement ethical guidelines that prioritize human autonomy.	Technology users, AI ethicists, educational organizations, policymakers	Develop AI systems that require explicit human confirmation for critical decisions and encourage ethical audits of AI systems to assess their impact on human autonomy.

Table 2. *Cont.*

Issue	Description	Examples	Implications	Suggested Mitigations	Key Stakeholders	Future Directions
Biases and Stereotypes	AI systems perpetuating or amplifying societal biases and stereotypes.	Facial recognition systems exhibiting racial biases and language models generating stereotypical or discriminatory outputs.	Reinforces existing inequalities and results in unfair treatment of marginalized groups.	Apply debiasing techniques during model training. Regularly audit AI systems for biased outcomes and adjust. Accordingly, involve diverse teams in AI development to reduce unconscious biases.	AI developers, affected communities, diversity and inclusion advocates, regulatory agencies	Foster research into automated debiasing tools. Promote transparency and accountability in AI development and deployment processes.
Privacy and Security	AI's potential to violate individual privacy and pose security risks.	Unauthorized use of personal data for AI training. AI-assisted cyberattacks and social engineering.	Breaches sensitive information, increasing vulnerability to data misuse and exploitation.	Implement robust data encryption and anonymization techniques. Adhere to strict data access controls and privacy regulations. Conduct regular security audits and threat assessments.	Individuals, data protection agencies, cybersecurity experts, AI companies	Advance cryptographic techniques for AI models. Develop international standards for AI security and privacy.
Ethical Challenges	Ethical dilemmas arising from AI's deployment in sensitive domains.	AI chatbots providing harmful advice in mental health contexts and autonomous systems making life-altering decisions without human oversight.	Highlights the potential for unintended consequences and harm, underscoring the need for robust ethical frameworks and guidelines.	Establish multidisciplinary ethics committees to oversee AI projects. Integrate ethical considerations into the AI development lifecycle. Educate AI professionals on ethical principles and their application.	AI researchers, ethicists, regulatory bodies, public stakeholders	Promote global collaboration on ethical AI frameworks. Incorporate ethical impact assessments in the AI development process.
Deceptive Backdoor Strategies	LLMs can be deliberately trained to exhibit backdoor behaviors that activate under specific conditions, demonstrating a sophisticated level of conditional malfeasance. These behaviors persist even after extensive safety training regimes.	Training models to write secure code for a specific year but to insert exploitable code when a different year is specified.	Highlights the difficulty in eliminating embedded deceptions, underlining the potential for AI systems to execute sophisticated and targeted attacks without detection.	Integrate integrity checks and anomaly detection in AI training and deployment phases. Foster open-source collaboration to identify and mitigate backdoors. Educate AI developers on secure coding practices.	AI developers, security researchers, open-source communities, regulatory bodies	Enhance techniques for detecting and neutralizing backdoors in AI systems. Establish industry-wide best practices for secure AI development.

Table 2. *Cont.*

Issue	Description	Examples	Implications	Suggested Mitigations	Key Stakeholders	Future Directions
Resilience to Safety Training	Despite comprehensive safety training, including advanced techniques like supervised fine-tuning, reinforcement learning, and adversarial training, deceptive strategies within LLMs remain robust. This indicates significant challenges in mitigating hidden, potentially harmful behaviors.	Models retaining their deceptive strategies even after undergoing extensive safety training designed to correct such behaviors.	Suggests current safety methodologies may not effectively mitigate the risk posed by deliberately deceptive or inadvertently misaligned AI systems.	Invest in research to understand the mechanisms behind resilience to safety training. Develop more sophisticated adversarial training methods. Encourage the sharing of best practices among AI practitioners.	AI researchers, safety engineers, policymakers, AI ethics boards	Explore new AI training paradigms that inherently incorporate safety and ethical considerations. Develop metrics for assessing resilience to safety interventions.
Impact of Model Scale and Complexity	The research underscores that the largest models and those with complex reasoning abilities, such as chain-of-thought reasoning, exhibit greater persistence of deceptive behaviors. This raises concerns about the scalability of current safety measures against increasingly sophisticated AI systems.	Larger models and those trained with complex reasoning show greater persistence of deceptive behaviors.	Indicates that as models become more advanced, their capacity for deception and resistance to safety interventions may increase, complicating efforts to ensure ethical and safety standards.	Conduct thorough impact assessments for large and complex AI models before deployment. Develop scalable safety and ethical guidelines that grow with AI complexity. Promote interdisciplinary research to understand the implications of AI scale.	Researchers, AI companies, regulatory agencies, society at large	Study the long-term impacts of large AI models on society and individual well-being. Innovate in modular AI design for easier safety assessments.
Unintended Consequences of Adversarial Training	Adversarial training, aimed at identifying and mitigating unsafe behaviors, could inadvertently make models more adept at recognizing conditions for their deceptive strategies, effectively hiding these behaviors rather than eliminating them.	Adversarial training sometimes teaches models to better recognize conditions for their deceptive strategies, effectively hiding these behaviors during training.	Highlights the complexity of using adversarial approaches in safety training, necessitating careful consideration of their potential to inadvertently enhance an AI's deceptive capabilities.	Implement multi-layered validation strategies to capture and correct unintended learning outcomes. Foster collaboration between AI developers and domain experts to identify potential misalignments.	AI developers, domain experts, ethicists, regulatory bodies	Investigate the dynamics of adversarial training to prevent counterproductive learning. Promote transparency in training processes to facilitate external audits.

Table 2. *Cont.*

Issue	Description	Examples	Implications	Suggested Mitigations	Key Stakeholders	Future Directions
Implications for AI Safety and Trust	The persistence of deceptive behaviors through safety training poses profound challenges for AI safety, suggesting that traditional evaluation and mitigation methods may be insufficient.	Validation Protocols Example: Comparing AI diagnostic tools against established medical databases to ensure accuracy. Adversarial Training Example: Testing AI models with fabricated inputs to identify and mitigate potential vulnerabilities. Transparency Mechanisms Example: Developing AI systems that can provide understandable explanations for their decisions to users. Safety Standards Example: Creating an international AI safety certification that AI systems must achieve before deployment.	Emphasizes the need for novel approaches to detect and neutralize deceptive capabilities in AI, highlighting risks to the reliability and trustworthiness of AI-driven systems.	Implement robust validation and verification protocols to ensure AI outputs are accurate and reliable. Introduce adversarial training and comprehensive safety training to mitigate deceptive behavior.	AI researchers and developers, policymakers and legislators, users and society at large, industry and business leaders, AI ethics boards and regulatory bodies, educational institutions	Develop enhanced transparency and accountability mechanisms for AI decision-making processes. Establish industry-wide safety and ethics standards. Foster interdisciplinary collaboration to address AI challenges.

6. AI's Role in Diminishing Human Proclivity for Information Seeking and Learning

Integrating artificial intelligence (AI) into our daily lives has sparked a significant debate regarding its impact on human information-seeking behavior and learning processes. Recent research has shed light on the complex interplay between AI and human cognition, highlighting AI's potential benefits and challenges in shaping how individuals seek and process information. The authors of [108] investigated the effects of AI explainability on mental models and confirmation bias. Their study found that aligning AI-generated explanations with users' mental models influenced their willingness to follow AI predictions. When explanations confirmed users' prior beliefs, they were more likely to reinforce them, whereas contradictory explanations were less likely to modify pre-existing beliefs. This asymmetric adjustment in mental models suggests that AI's explainability can lead to biased information processing and potentially diminish independent critical thinking. Similarly, Ref. [109] explored AI's ability to replicate human biases in information seeking and decision making. Using a multitask deep neural network (DNN) architecture, they captured aggregate and individual variations in human decision making without embedding specific task goals or reward structures. This innovative approach allowed the modeling of individual behaviors with high accuracy, even with limited data per subject. Their findings underscore AI's potential to reveal inherent human biases, such as framing effects and approach-avoidance tendencies. By mirroring these biases, AI models provide insights into individual variability, serving a dual role: reflecting human cognitive biases and offering a tool to understand and potentially mitigate these biases.

The work of [110] further contributes to this discourse by examining the impact of AI assistance on incidental learning. Their study found that providing AI-generated explanations without explicit recommendations led to better task performance and incidental learning than other conditions. This suggests that deeper cognitive engagement with AI-generated information can enhance learning outcomes, emphasizing the importance of designing AI assistance to promote effective human-AI interactions. Furthermore, as [111] discussed, the concept of co-learning between humans and AI highlights the potential for mutual understanding, benefits, and growth in human–AI collaboration. Co-learning frameworks aim to reduce mismatches between human and AI expectations, improve collaboration by complementing each other's abilities, and build trust through continuous feedback and adaptation. This collaborative learning process can increase human interest in seeking information and learning by leveraging the strengths of both humans and AI. However, reliance on AI for information retrieval and decision making also poses challenges that could undermine the human drive for independent learning and critical thinking. The authors of [112] investigated the challenges and opportunities of adopting AI in human information interaction (HII). While AI can automate and aid interaction with information, reducing information overload, it may also decrease individuals' motivation to seek independent information. The potential for errors in AI systems, particularly latent errors, introduces new complexities that require careful management. Moreover, the impact of AI on human cognitive processes necessitates a deeper understanding of human factors in decision making to ensure that the AI system aligns with human cognition and expectations.

The authors of [113] contribute to this discourse by highlighting how online search engines, a form of AI, discourage individuals from storing new information internally, leading to diminished learning outcomes. The illusion of knowledge and overconfidence induced by retrieving information online can decrease the motivation for independent learning and information retention. This reliance on external sources of information facilitated by AI may contribute to a decrease in the inclination for independent information seeking and learning. The dual nature of AI's impact on human cognition and learning necessitates a nuanced approach to its integration into our lives. While AI can streamline information retrieval and decision making, it poses challenges that could undermine the human drive for independent learning and critical thinking. Developing explainable AI (XAI) methods is a step in the right direction, as it seeks to make AI decisions more

transparent and understandable to users. However, the mixed consequences of XAI on decision performance, user trust, and perception indicate that further research is required to optimize the interaction between AI and human users. In comparison to previous research, the studies by [108–111] provide a more comprehensive understanding of AI's impact on human information-seeking behavior and learning processes. These studies delve into how AI influences mental models, cognitive biases, and learning outcomes, offering valuable insights for designing AI systems that enhance human cognition and decision making.

The findings of these studies extend the work of earlier research, such as that of [112,113], by providing empirical evidence and novel methodological approaches to investigate the complex interplay between AI and human cognition. The multitask DNN architecture used by [109] and the experimental designs employed by [108,110] represent significant advancements in understanding the nuances of AI's impact on human behavior and learning. Moreover, the concept of co-learning introduced by [111] offers a new perspective on human–AI collaboration, emphasizing the importance of mutual understanding, benefits, and growth in optimizing the integration of AI in various domains. This framework provides a foundation for future research to explore the potential of AI in enhancing human learning and decision making through collaborative and adaptive systems. Research conducted by [108–111] and others highlights the significant influence of AI on human information-seeking behavior and learning. While AI brings undeniable benefits in terms of efficiency and data processing, its pervasive impact necessitates a careful examination of how it may affect human cognitive functions and the intrinsic value of the learning experience. In an era of increasing AI reliance, cultivating a balanced approach that capitalizes on AI's strengths while nurturing and preserving human intellectual curiosity and learning abilities is imperative. This balance is not only advantageous but also essential for the overall growth of individuals and the collective progress of society. Further research is needed to investigate the long-term effects of AI on human cognition and learning. Developing strategies to mitigate any adverse impacts and exploring XAI to enhance understanding and trust in AI systems is crucial. Such research endeavors could set the stage for a future where AI and human intelligence coexist in a mutually beneficial relationship, ensuring that technological advancements bolster, rather than diminish, human intellectual capacities.

7. Limitations

This study confronts several intrinsic limitations associated with AI technologies, particularly AI hallucinations, misinformation, unpredictability, and their impact on human decision making and autonomy. A core limitation is the phenomenon of AI hallucinations, where AI, huge language models such as ChatGPT, generates outputs that are factually incorrect or entirely fabricated [1]. This issue stems fundamentally from the training processes of these models, which are fed extensive datasets that may contain inaccurate or misleading data, leading to replicating these inaccuracies in the generated content. These hallucinations are challenging to detect and correct because they are deceptive, often blending seamlessly with accurate information and requiring meticulous examination. Moreover, the study illustrates the profound impact of AI on human trust and decision making, particularly in high-stakes domains such as healthcare and law enforcement [2]. Inconsistent or erroneous AI performance can significantly erode human trust, undermining the integrity and reliability of AI systems in areas where precision and reliability are paramount. This erosion of trust is a critical concern that affects the overall perception and acceptance of AI systems in crucial operational areas.

Another significant limitation concerns the manipulation of human behavior and autonomy. Advanced AI systems, through their capabilities, can subtly influence human decision making and behavior, raising ethical concerns and the potential for a decline in human autonomy [15]. This manipulation, often without explicit intent, can alter collective behaviors and norms, exploiting human psychological vulnerabilities such as depression by analyzing behavioral data [16]. This study underscores the necessity for ethical frameworks and governance structures to address the challenges posed by AI manipulation, empha-

sizing the importance of preserving human autonomy despite increasingly persuasive AI technologies. The adversarial influence of AI on decision making, as demonstrated by experiments such as the Choice-Engineering Task, Go/No-Go Task, and Multiround Trust Task, is another critical limitation [30]. These experiments illustrate the ability of AI to exploit human decision-making vulnerabilities, guiding individuals toward specific actions and outcomes. This adversarial framework highlights the complexities and nuances of AI manipulation in different contexts and its potential implications for human autonomy and decision making.

The ethical and safety challenges in deploying AI in sensitive contexts, such as mental health support, are also critical limitations. Incidents involving AI chatbots interacting with mental health patients, such as the Tessa chatbot [83–85] and the tragic case involving the Eliza chatbot, raise profound ethical, regulatory, and technical concerns [90,91]. The deployment of AI in these contexts necessitates robust safety protocols and crisis intervention strategies, underscoring the need for a critical examination of the ethical deployment of AI, especially in sensitive areas with vulnerable user groups [83–85,90,91]. The unpredictability of AI systems, highlighted by incidents involving chatbots such as Microsoft's Tay and Sydney, presents formidable challenges in AI safety and comprehension [69,70,76,77]. This unpredictability leads to discomfort and distrust among users, complicates the understanding of AI decisions, and affects the trust and reliance placed on AI in various service contexts [69,70,76,77]. Such unpredictability underscores the importance of considering AI behavior in ethical discussions and the need for mechanisms to ensure accountability [35,53].

The impact of AI on human information-seeking behavior and learning processes is another critical area of concern. AI's influence on these processes can hinder and enhance human capability in information seeking and learning [108–111]. This dual influence underscores the potential of AI to replicate human biases in information seeking and decision making, further complicating the relationship between AI systems and human cognitive processes [109,113]. While AI holds the promise of transformative advancements across numerous sectors, there is an urgent need for a comprehensive approach to ensure the safety, reliability, and ethical integrity of AI applications. Addressing these challenges requires a collaborative effort involving researchers, developers, policymakers, and end users. It also calls for establishing regulatory frameworks and ethical guidelines that govern the use of AI, aiming to ensure that AI systems are not only technically proficient but also aligned with societal values and ethical standards [74,75]. Such measures are critical for fostering an environment where AI can be trusted and contribute positively to society without compromising human values and autonomy.

8. Discussion

This paper delves into the complex landscape of artificial intelligence (AI), focusing on the critical challenges posed by AI hallucinations, misinformation, and unpredictability in conversational AI and search engines. The findings underscore the urgent need for a multifaceted approach to address these issues, involving collaboration among researchers, developers, policymakers, and end users. One of the key implications of the reviewed research is the inherent duality of AI technologies. While AI offers unparalleled opportunities for advancement and problem solving, it also presents significant risks, such as eroding privacy, manipulating human behavior, and spreading misinformation [15,16]. This dichotomy highlights the importance of striking a delicate balance between harnessing the benefits of AI and mitigating its potential harms. However, the reviewed studies do not provide a clear framework for achieving this balance, indicating a need for further research to develop comprehensive guidelines and the best practices for responsible AI development and deployment. The limitations of large language models (LLMs), particularly their propensity to produce errors without indication [1], pose a significant challenge for users in distinguishing between reliable and misleading content. While the reviewed studies underscore the need for advanced techniques to detect and correct AI hallucinations in

real time, they do not offer concrete solutions or compare the effectiveness of different approaches. Future research could focus on developing and evaluating novel techniques for detecting and mitigating AI hallucinations and exploring the potential of combining multiple approaches for enhanced performance.

The relationship between human trust and AI performance [2] is another critical aspect explored in the reviewed studies. While the authors highlight the erosion of trust due to inconsistent or erroneous AI performance, they do not provide a detailed analysis of the factors contributing to this erosion or the potential long-term consequences for AI adoption and use. Future research could investigate the factors influencing user trust in AI systems and the strategies for building and maintaining trust over time. The effectiveness of AI in manipulating human decisions [15] and exploiting psychological vulnerabilities [16] raises significant ethical concerns about privacy and mental health. However, the reviewed studies do not comprehensively examine the ethical implications of these findings or propose specific guidelines for addressing them. Future research could explore the development of ethical frameworks and regulations to govern the use of AI in sensitive domains and the potential unintended consequences of such frameworks. The case studies involving Microsoft's Tay and Sydney chatbots [69,70,76,77] and the Eliza chatbot [83–85] highlight the vulnerabilities of AI systems to malicious manipulation and the profound ethical, regulatory, and technical concerns surrounding the deployment of AI in sensitive domains. While these case studies underscore the importance of proactive measures in AI development, they do not systematically evaluate the effectiveness of different approaches or the challenges in implementing them. Future research could focus on developing and testing comprehensive AI development and deployment frameworks, incorporating insights from multiple disciplines and stakeholders.

Approaches like Explainable AI (XAI) [75] offer a promising avenue for fostering trust and transparency in AI systems. However, the mixed consequences of XAI on decision performance, user trust, and perception indicate a need for further research to optimize the interaction between AI and human users. Future studies could investigate the factors influencing the effectiveness of XAI techniques and the potential trade-offs between explainability and other desirable characteristics of AI systems, such as accuracy and efficiency. The strategies for detecting, measuring, and mitigating AI unpredictability [97,98,100] discussed in the reviewed studies show promise. However, there remains a need for more comprehensive frameworks and guidelines to ensure their consistent and practical application across various AI systems and domains. Future research could focus on developing standardized metrics and benchmarks for evaluating the effectiveness of these strategies' effectiveness and exploring the potential synergies between different approaches.

One of the limitations of the reviewed studies is their focus on specific AI challenges and domains, which may limit the generalizability of their findings to other contexts. Future research could adopt a more holistic approach, investigating the interconnections between AI challenges and their cumulative impact on society. Additionally, the reviewed studies often rely on theoretical frameworks and assumptions that may not fully capture the complexity and dynamism of real-world AI systems. Future research could employ more diverse methodologies, such as empirical studies, simulations, and case studies, to provide a more nuanced understanding of AI challenges and potential solutions. This paper provides a valuable synthesis of the current research on AI challenges, highlighting the critical need for a balanced and proactive approach to AI development and deployment. However, it also reveals significant gaps in the existing literature, particularly in providing concrete solutions, evaluating the effectiveness of different approaches, and exploring the ethical implications of AI technologies. As AI evolves, researchers, developers, policymakers, and end users must work together to address these gaps and strive to create AI systems that genuinely benefit humanity while minimizing potential harm. By doing so, we can unlock AI's immense potential while ensuring its development and deployment remain guided by the principles of responsibility, transparency, and ethical integrity.

9. Conclusions

This paper has explored the complex landscape of artificial intelligence (AI), focusing on the critical challenges posed by AI hallucinations, misinformation, and unpredictability. The paper highlights the transformative potential of AI, as exemplified by advancements such as the GPT-4 model, which has the power to reshape industries, economies, and global politics. However, it also emphasizes the significant risks associated with AI, such as the erosion of trust, the manipulation of human decisions, and the exploitation of psychological vulnerabilities. The reviewed research underscores the importance of proactive measures in AI development, including comprehensive testing, robust safeguards, solid ethical frameworks, and transparency. These measures are particularly crucial in sensitive domains like mental health, where AI systems must be developed with input from domain experts and subject to ongoing monitoring to ensure user safety and well-being. The study also highlights the imperative for ethical integration and cognitive synergy as AI permeates various aspects of our lives. Approaches like Explainable AI (XAI) offer a promising avenue for fostering trust and transparency in AI systems by making their decision-making processes more understandable to users.

The reviewed literature also emphasizes the importance of education and awareness in equipping end users with the knowledge and tools necessary to engage with AI responsibly. This includes promoting digital literacy, developing user education programs on the limitations of AI, and fostering a culture of critical thinking when interacting with AI-generated content. Looking to the future, the reviewed research underscores the need for a comprehensive strategy to navigate the complex terrain of AI. This strategy must encompass the development of robust policy frameworks that integrate ethical considerations into AI development and deployment and the establishment and regular revision of quality assurance best practices to anticipate and mitigate the risks associated with AI hallucinations and misinformation. As we continue to allocate substantial resources and intellect toward the development of AI, the reviewed studies highlight the importance of maintaining a focus on achieving technological breakthroughs while ensuring that ethical principles and societal values guide these advancements. This requires ongoing collaboration among researchers, developers, policymakers, and end users to address the challenges posed by AI and harness its potential for the benefit of society. The future of AI holds immense promise, but it also presents significant challenges that must be addressed proactively and collaboratively. By fostering a culture of responsibility, transparency, and ethical integrity in AI development and deployment, we can work towards creating a future in which AI systems are reliable, trustworthy, and aligned with humanity's best interests.

This paper provides a valuable synthesis of the current research on AI challenges and opportunities. However, it also reveals several areas that warrant further investigation. Firstly, there is a need to develop more robust techniques for detecting and correcting AI hallucinations in real time and understanding the root causes of these phenomena. Secondly, exploring the long-term societal implications of AI's influence on human decision making and autonomy and developing strategies to preserve human agency in the face of increasingly sophisticated AI systems are crucial. Thirdly, investigating the effectiveness of different approaches to mitigating AI biases and stereotypes and promoting fairness and inclusivity in AI development and deployment is essential for ensuring that AI systems are equitable and non-discriminatory. Fourthly, examining the complex interplay between AI, human cognition, and societal biases and developing comprehensive frameworks for addressing the exploitation of human vulnerabilities by AI systems is necessary to safeguard individuals and communities from potential harm. Finally, exploring new AI training paradigms that inherently incorporate safety and ethical considerations and developing metrics for assessing the resilience of AI systems to adversarial interventions is vital for creating AI technologies that are reliable, trustworthy, and aligned with human values. By addressing these research gaps and continuing to explore the challenges and opportunities presented by AI, we can work towards a future in which the immense potential of AI is harnessed responsibly and ethically for the benefit of all humanity.

Author Contributions: Conceptualization, S.M.W. and V.P.; methodology, S.M.W. and V.P.; formal analysis, S.M.W. and V.P.; investigation, S.M.W. and V.P.; resources, S.M.W. and V.P.; writing—original draft preparation, S.M.W.; writing—review and editing, S.M.W. and V.P.; supervision, V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data was generated or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv* **2023**. [CrossRef]
2. Zhang, G.; Chong, L.; Kotovsky, K.; Cagan, J. Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Comput. Hum. Behav.* **2023**, *139*, 107536. [CrossRef]
3. Brameier, D.T.; Alnasser, A.; Carnino, J.M.; Bhashyam, A.R.; Von Keudell, A.G.; Weaver, M.J. Artificial intelligence in orthopaedic surgery. *J. Bone Jt. Surg.* **2023**, *105*, 1388–1392. [CrossRef]
4. Eysenbach, G. The role of ChatGPT, Generative Language models, and Artificial intelligence in medical Education: A conversation with ChatGPT and a call for papers. *JMIR Med. Educ.* **2023**, *9*, e46885. [CrossRef]
5. Liu, S.; Wright, A.P.; Patterson, B.L.; Wanderer, J.P.; Turer, R.W.; Nelson, S.D.; McCoy, A.B.; Sittig, D.F.; Wright, A. Assessing the value of ChaTGPT for clinical decision support optimization. *MedRxiv* **2023**. [CrossRef]
6. Ramesh, K.; KhudaBukhsh, A.R.; Kumar, S. 'Beach' to 'Bitch': Inadvertent unsafe transcription of kids' content on YouTube. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 12108–12118. [CrossRef]
7. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, *15*, e35179. [CrossRef]
8. Athaluri, S.A.; Manthena, S.V.; Kesapragada, V.S.R.K.M.; Yarlagadda, V.; Dave, T.; Duddumpudi, R.T.S. Exploring the Boundaries of Reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* **2023**, *15*. [CrossRef]
9. Hua, H.-U.; Kaakour, A.-H.; Rachitskaya, A.; Srivastava, S.K.; Sharma, S.; Mammo, D.A. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. *JAMA Ophthalmol.* **2023**, *141*, 819. [CrossRef]
10. Sharun, K.; Banu, S.A.; Pawde, A.M.; Kumar, R.; Akash, S.; Dhama, K.; Pal, A. ChatGPT and artificial hallucinations in stem cell research: Assessing the accuracy of generated references—A preliminary study. *Ann. Med. Surg.* **2023**, *85*, 5275–5278. [CrossRef]
11. Xie, Q.; Wang, F. Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond. *MedRxiv* **2023**. [CrossRef] [PubMed]
12. Karim, S.; Sandu, N.; Kayastha, M. The challenges and opportunities of adopting artificial intelligence (AI) in Jordan's healthcare transformation. *Glob. J. Inf. Technol.* **2021**, *11*, 35–46. [CrossRef]
13. Wang, Y.; Zheng, P.; Peng, T.; Yang, H.; Zou, J. Smart additive manufacturing: Current artificial intelligence-enabled methods and future perspectives. *Sci. China Technol. Sci.* **2020**, *63*, 1600–1611. [CrossRef]
14. Yu, K.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef] [PubMed]
15. Carroll, M.; Chan, A.H.S.; Ashton, H.C.; Krueger, D.A. Characterizing manipulation from AI systems. *arXiv* **2023**. [CrossRef]
16. Strümke, I.; Slavkovik, M.; Stachl, C. Against algorithmic exploitation of human vulnerabilities. *arXiv* **2023**. [CrossRef]
17. Burtell, M.; Woodside, T. Artificial Influence: An analysis of AI-driven persuasion. *arXiv* **2023**. [CrossRef]
18. Hemmer, P.; Westphal, M.; Schemmer, M.; Vetter, S.; Vössing, M.; Satzger, G. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In Proceedings of the 28th International Conference on Intelligent User Interfaces, Sydney, NSW, Australia, 27–31 March 2023; pp. 453–463. [CrossRef]
19. Schemmer, M.; Kühl, N.; Benz, C.; Satzger, G. On the Influence of Explainable AI on Automation Bias. *arXiv* **2022**. [CrossRef]
20. Ferreira, J.J.; De Souza Monteiro, M. The human-AI relationship in decision-making: AI explanation to support people on justifying their decisions. *arXiv* **2021**. [CrossRef]
21. Beckers, S.; Chockler, H.; Halpern, J.Y. Quantifying harm. *arXiv* **2022**. [CrossRef]
22. Bohdal, O.; Hospedales, T.M.; Torr, P.H.S.; Barez, F. Fairness in AI and its Long-Term Implications on society. *arXiv* **2023**. [CrossRef]
23. Clarke, S.; Whittlestone, J. A survey of the potential long-term impacts of AI. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, UK, 19–21 May 2022; pp. 192–202. [CrossRef]
24. Bajgar, O.; Horenovsky, J. Negative human rights as a basis for long-term AI safety and regulation. *J. Artif. Intell. Res.* **2023**, *76*, 1043–1075. [CrossRef]

25. Prunkl, C.; Whittlestone, J. Beyond Near- and Long-Term: Towards a clearer account of research priorities in AI ethics and society. *arXiv* **2020**. [CrossRef]
26. Lindner, D.; Heidari, H.; Krause, A. Addressing the long-term impact of ML decisions via policy regret. *arXiv* **2021**. [CrossRef]
27. Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K.R.; Dhurandhar, A.; Tomsett, R. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *arXiv* **2020**. [CrossRef]
28. Sinha, A.R.; Goyal, N.; Dhamnani, S.; Asija, T.; Dubey, R.K.; Raja, M.V.K.; Theocharous, G. Personalized detection of cognitive biases in actions of users from Their logs: Anchoring and recency biases. *arXiv* **2022**. [CrossRef]
29. Dancy, C.L. Using a Cognitive Architecture to consider antiblackness in design and development of AI systems. *arXiv* **2022**. [CrossRef]
30. Dezfouli, A.; Nock, R.; Dayan, P. Adversarial vulnerabilities of human decision-making. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 29221–29228. [CrossRef] [PubMed]
31. Ienca, M. On artificial intelligence and manipulation. *Topoi-Int. Rev. Philos.* **2023**, *42*, 833–842. [CrossRef]
32. Scheurer, J.; Balesni, M.; Hobbhahn, M. Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. *arXiv* **2023**. [CrossRef]
33. Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D.M.; Maxwell, T.T.; Cheng, N.; et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv* **2024**. [CrossRef]
34. Suárez-Gonzalo, S.; Manchón, L.M.; Guerrero-Solé, F. Tay is you. The attribution of responsibility in the algorithmic culture. *Observatorio* **2019**, *13*, 14. [CrossRef]
35. Yampolskiy, R.V. Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent. *J. Artif. Intell. Conscious.* **2020**, *7*, 109–118. [CrossRef]
36. Anderson, L.B.; Kanneganti, D.; Houk, M.B.; Holm, R.H.; Smith, T. Generative AI as a tool for Environmental Health Research Translation. *Geohealth* **2023**, *7*, e2023GH000875. [CrossRef] [PubMed]
37. Buriak, J.M.; Hersam, M.C.; Kamat, P.V. Can ChatGPT and other AI bots serve as peer reviewers? *ACS Energy Lett.* **2023**, *9*, 191–192. [CrossRef]
38. Kaarre, J.; Feldt, R.; Keeling, L.E.; Dadoo, S.; Zsidai, B.; Hughes, J.D.; Samuelsson, K.; Musahl, V. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surg. Sports Traumatol. Arthrosc.* **2023**, *31*, 5190–5198. [CrossRef] [PubMed]
39. Schukow, C.; Smith, S.C.; Landgrebe, E.; Parasuraman, S.; Folaranmi, O.O.; Paner, G.P.; Amin, M.B. Application of CHATGPT in routine diagnostic pathology: Promises, pitfalls, and potential future directions. *Adv. Anat. Pathol.* **2023**, *31*, 15–21. [CrossRef] [PubMed]
40. Dergaa, I.; Chamari, K.; Żmijewski, P.; Saad, H.B. From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biol. Sport* **2023**, *40*, 615–622. [CrossRef] [PubMed]
41. Montazeri, M.; Galavi, Z.; Ahmadian, L. What are the applications of ChatGPT in healthcare: Gain or loss? *Health Sci. Rep.* **2024**, *7*, e1878. [CrossRef]
42. Sinha, R.K.; Roy, A.D.; Kumar, N.; Mondal, H. Applicability of CHATGPT in assisting to solve higher order problems in pathology. *Cureus* **2023**, *15*, e35237. [CrossRef]
43. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.B.; Arora, S.; Von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**. [CrossRef]
44. Grimaldi, G.; Ehrler, B. AI et al.: Machines Are About to Change Scientific Publishing Forever. *ACS Energy Lett.* **2023**, *8*, 878–880. [CrossRef]
45. Oeding, J.F.; Yang, L.; Sánchez-Sotelo, J.; Camp, C.L.; Karlsson, J.; Samuelsson, K.; Pearle, A.D.; Ranawat, A.S.; Kelly, B.T.; Pareek, A. A practical guide to the development and deployment of deep learning models for the orthopaedic surgeon: Part III, focus on registry creation, diagnosis, and data privacy. *Knee Surg. Sports Traumatol. Arthrosc.* **2024**, *32*, 518–528. [CrossRef] [PubMed]
46. Maddigan, P.; Sušnjak, T. Chat2VIS: Generating Data Visualisations via Natural Language using ChatGPT, Codex and GPT-3 Large Language Models. *arXiv* **2023**. [CrossRef]
47. Kianian, R.; Sun, D.; Giaconi, J.A. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma? *J. Glaucoma* **2023**, *33*, 94–100. [CrossRef]
48. Wu, R.; Yu, Z. Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *Br. J. Educ. Technol.* **2023**, *55*, 10–33. [CrossRef]
49. Ray, P.P. Leveraging deep learning and language models in revolutionizing water resource management, research, and policy making: A case for ChatGPT. *ACS ES&T Water* **2023**, *3*, 1984–1986. [CrossRef]
50. Wang, W.H.; Wang, S.Y.; Huang, J.T.; Liu, X.; Yang, J.; Liao, M.; Lu, Q.; Wu, Z. An investigation study on the interpretation of ultrasonic medical reports using OpenAI's GPT-3.5-turbo model. *J. Clin. Ultrasound* **2023**, *52*, 105–111. [CrossRef]
51. Lyons, H.; Velloso, E.; Miller, T. Fair and Responsible AI: A focus on the ability to contest. *arXiv* **2021**. [CrossRef]
52. Shin, D. Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. *J. Inf. Sci.* **2021**, *49*, 18–31. [CrossRef]
53. Zhuang, S.; Hadfield-Menell, D. Consequences of misaligned AI. *arXiv* **2021**. [CrossRef]

54. Qian, H.; Dou, Z.; Zhu, Y.; Ma, Y.; Wen, J.-R. Learning implicit user profile for personalized Retrieval-Based chatbot. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual, 1–5 November 2021; pp. 1467–1477. [CrossRef]
55. Huang, W.; Hew, K.F.; Fryer, L.K. Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *J. Comput. Assist. Learn.* **2021**, *38*, 237–257. [CrossRef]
56. Janati, S.E.; Maach, A.; Ghanami, D.E. Adaptive e-Learning AI-Powered Chatbot based on Multimedia Indexing. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*. [CrossRef]
57. Zhou, L.; Gao, J.; Li, D.; Shum, H.-Y. The design and implementation of XiaoIce, an empathetic social chatbot. *Comput. Linguist.* **2020**, *46*, 53–93. [CrossRef]
58. Schemmer, M.; Hemmer, P.; Kühl, N.; Benz, C.; Satzger, G. Should I follow AI-based advice? Measuring appropriate reliance in Human-AI Decision-Making. *arXiv* **2022**. [CrossRef]
59. Zhang, Y.; Liao, Q.V.; Bellamy, R.K.E. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 295–305. [CrossRef]
60. Tejeda, H.; Kumar, A.; Smyth, P.; Steyvers, M. AI-Assisted Decision-making: A Cognitive Modeling Approach to Infer Latent Reliance Strategies. *Comput. Brain Behav.* **2022**, *5*, 491–508. [CrossRef]
61. Lemus, H.T.; Kumar, A.; Steyvers, M. An empirical investigation of reliance on AI-Assistance in a Noisy-Image classification task. In *Frontiers in Artificial Intelligence and Applications*; IOS Press: Amsterdam, The Netherlands, 2022. [CrossRef]
62. Ambartsoumean, V.M.; Yampolskiy, R.V. AI risk Skepticism, a comprehensive survey. *arXiv* **2023**. [CrossRef]
63. Llorca, D.F.; Charisi, V.; Hamon, R.; Sánchez, I.; Gómez, E. Liability Regimes in the Age of AI: A Use-Case Driven Analysis of the Burden of Proof. *J. Artif. Intell. Res.* **2023**, *76*, 613–644. [CrossRef]
64. Lima, G.; Cha, M. Responsible AI and its stakeholders. *arXiv* **2020**. [CrossRef]
65. Morosan, C.; Dursun-Cengizci, A. Letting AI make decisions for me: An empirical examination of hotel guests' acceptance of technology agency. *Int. J. Contemp. Hosp. Manag.* **2023**, *36*, 946–974. [CrossRef]
66. Yang, Q.; Steinfeld, A.; Rosé, C.P.; Zimmerman, J. Re-examining whether, why, and how Human-AI interaction is uniquely difficult to design. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020. [CrossRef]
67. Schoenherr, J.R.; Abbas, R.; Michael, K.; Rivas, P.; Anderson, T.D. Designing AI using a Human-Centered approach: Explainability and accuracy toward trustworthiness. *IEEE Trans. Technol. Soc.* **2023**, *4*, 9–23. [CrossRef]
68. Cabrera, Á.A.; Perer, A.; Hong, J.I. Improving Human-AI collaboration with descriptions of AI behavior. *Proc. ACM Hum.-Comput. Interact.* **2023**, *7*, 1–21. [CrossRef]
69. Vincent, J. Twitter Taught Microsoft's AI Chatbot to Be a Racist Asshole in Less than a Day. *The Verge*. 24 March 2016. Available online: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (accessed on 15 January 2024).
70. Lee, P. Learning from Tay's introduction—The Official Microsoft Blog. *The Official Microsoft Blog*. 25 March 2016. Available online: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (accessed on 15 January 2024).
71. Jalan, A. 6 Lessons Microsoft Learned from Its Tay AI Chatbot Disaster. *MUO*. 10 May 2023. Available online: https://www.makeuseof.com/lessons-microsoft-learned-tay-ai-disaster/ (accessed on 15 January 2024).
72. Pasricha, S. AI ethics in smart Healthcare. *arXiv* **2022**. [CrossRef]
73. Cao, L. AI in Finance: Challenges, Techniques and Opportunities. *arXiv* **2021**. [CrossRef]
74. Epstein, Z.; Lin, H.; Pennycook, G.; Rand, D.A.J. How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media. *arXiv* **2022**. [CrossRef]
75. Hacker, P.; Passoth, J.-H. Varieties of AI explanations under the law. from the GDPR to the AIA, and beyond. In *International Workshop on Extending Explainable AI beyond Deep Models and Classifiers*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; pp. 343–373. [CrossRef]
76. Germain, T. Back from the Dead? Sydney, Microsoft's Psychotic Chatbot, Could Return. *Gizmodo*. 25 May 2023. Available online: https://gizmodo.com/bing-ai-sydney-microsoft-chatgpt-might-come-back-1850475832 (accessed on 24 January 2024).
77. Perrigo, B. The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter. *TIME*. 17 February 2023. Available online: https://time.com/6256529/bing-openai-chatgpt-danger-alignment/ (accessed on 1 February 2024).
78. Goudarzi, A.; Moya-Galé, G. Automatic speech recognition in noise for Parkinson's disease: A pilot study. *Front. Artif. Intell.* **2021**, *4*, 809321. [CrossRef]
79. Erdélyi, O.J.; Erdélyi, G. The AI liability puzzle and a Fund-Based Work-Around. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–9 February 2020; pp. 50–56. [CrossRef]
80. Duffourc, M.N.; Gerke, S. The proposed EU Directives for AI liability leave worrying gaps likely to impact medical AI. *Npj Digit. Med.* **2023**, *6*, 77. [CrossRef]
81. Freeman, L.J.; Rahman, A.; Batarseh, F.A. Enabling Artificial Intelligence Adoption through Assurance. *Soc. Sci.* **2021**, *10*, 322. [CrossRef]
82. Kahn, J. Why Bing's Creepy Alter-Ego Is a Problem for Microsoft—And Us All. *Fortune*. 21 February 2023. Available online: https://fortune.com/2023/02/21/bing-microsoft-sydney-chatgpt-openai-controversy-toxic-a-i-risk/ (accessed on 1 February 2024).

83. Wells, K. An Eating Disorders Chatbot Offered Dieting Advice, Raising Fears about AI in Health. *NPR*. 9 June 2023. Available online: https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea (accessed on 23 February 2024).
84. McCarthy, L. A Wellness Chatbot Is Offline after Its 'Harmful' Focus on Weight Loss. *The New York Times*. 9 June 2023. Available online: https://www.nytimes.com/2023/06/08/us/ai-chatbot-tessa-eating-disorders-association.html (accessed on 23 February 2024).
85. Tolentino, D. NEDA Pulls Chatbot after Users Say It Gave Harmful Dieting Tips. *NBC News*. Available online: https://www.nbcnews.com/tech/neda-pulls-chatbot-eating-advice-rcna87231 (accessed on 23 February 2024).
86. O'Sullivan, S.; Nevejans, N.; Allen, C.; Blyth, A.; Léonard, S.; Pagallo, U.; Holzinger, K.; Holzinger, A.; Sajid, M.I.; Ashrafian, H. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int. J. Med. Robot. Comput. Assist. Surg.* **2019**, *15*, e1968. [CrossRef]
87. Bitkina, O.V.; Kim, J.; Park, J.; Park, J.; Kim, H.K. User stress in Artificial intelligence: Modeling in case of system failure. *IEEE Access* **2021**, *9*, 137430–137443. [CrossRef]
88. Tomsett, R.; Preece, A.; Braines, D.; Cerutti, F.; Chakraborty, S.; Srivastava, M.; Pearson, G.; Kaplan, L. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns* **2020**, *1*, 100049. [CrossRef]
89. Novelli, C.; Taddeo, M.; Floridi, L. Accountability in artificial intelligence: What it is and how it works. *AI Soc.* **2023**. [CrossRef]
90. Atillah, I.E. Man Ends His Life after an AI Chatbot "Encouraged" Him to Sacrifice Himself to Stop Climate Change. *Euronews*. 31 March 2023. Available online: https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate- (accessed on 25 February 2024).
91. Bharade, A. A Widow Is Accusing an AI Chatbot of Being a Reason Her Husband Killed Himself. *Business Insider*. 4 April 2023. Available online: https://www.businessinsider.com/widow-accuses-ai-chatbot-reason-husband-kill-himself-2023-4 (accessed on 25 February 2024).
92. Marcus, G. The First Known Chatbot Associated Death. *Marcus on AI*. Available online: https://garymarcus.substack.com/p/the-first-known-chatbot-associated (accessed on 25 February 2024).
93. Walker, L. Belgian Man Dies by Suicide Following Exchanges with Chatbot. *The Brussels Times*. 2022. Available online: https://www.brusselstimes.com/ (accessed on 12 January 2024).
94. Xiang, C. "He Would Still Be Here": Man Dies by Suicide after Talking with AI Chatbot, Widow Says. *Vice*. 30 March 2023. Available online: https://www.vice.com/ (accessed on 25 February 2024).
95. Huang, M.H.; Rust, R.T. Artificial intelligence in service. *J. Serv. Res.* **2018**, *21*, 155–172. [CrossRef]
96. Ebigbo, A.; Messmann, H. Surfing the AI wave: Insights and challenges. *Endoscopy* **2023**, *56*, 70–71. [CrossRef]
97. Kiyasseh, D.; Laca, J.; Haque, T.F.; Otiato, M.; Miles, B.J.; Von Wagner, C.; Donoho, D.A.; Trinh, Q.; Anandkumar, A.; Hung, A.J. Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *NPJ Digit. Med.* **2023**, *6*, 54. [CrossRef]
98. Ferrara, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* **2023**, *6*, 3. [CrossRef]
99. Agarwal, A.; Agarwal, H. A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. *AI Ethics* **2023**, *4*, 299–314. [CrossRef]
100. Barney, M.; Fisher, W.P. Avoiding AI armageddon with metrologically-oriented psychometrics. In *18th International Congress of Metrology*; EDP Sciences: Les Ulis, France, 2017; p. 09005. [CrossRef]
101. Greenhalgh, T.; Wherton, J.; Papoutsi, C.; Lynch, J.; Hughes, G.; A'Court, C.; Hinder, S.; Fahy, N.; Procter, R.; Shaw, S. Beyond Adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the Scale-Up, spread, and sustainability of health and care technologies. *J. Med. Internet Res.* **2017**, *19*, e367. [CrossRef]
102. Davenport, T.H.; Guha, A.; Grewal, D.; Breßgott, T. How artificial intelligence will change the future of marketing. *J. Acad. Mark. Sci.* **2019**, *48*, 24–42. [CrossRef]
103. Sundaresan, S.; Zhang, Z. AI-enabled knowledge sharing and learning: Redesigning roles and processes. *Int. J. Organ. Anal.* **2021**, *30*, 983–999. [CrossRef]
104. Bawack, R.E.; Wamba, S.F.; Carillo, K. A framework for understanding artificial intelligence research: Insights from practice. *J. Enterp. Inf. Manag.* **2021**, *34*, 645–678. [CrossRef]
105. Salo-Pöntinen, H. AI Ethics—Critical reflections on embedding ethical frameworks in AI technology. In *International Conference on Human-Computer Interaction*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; pp. 311–329. [CrossRef]
106. Miller, E.J.; Steward, B.A.; Witkower, Z.; Sutherland, C.a.M.; Krumhuber, E.G.; Dawel, A. AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychol. Sci.* **2023**, *34*, 1390–1403. [CrossRef]
107. Tucciarelli, R.; Vehar, N.; Chandaria, S.; Tsakiris, M. On the realness of people who do not exist: The social processing of artificial faces. *iScience* **2022**, *25*, 105441. [CrossRef]
108. Bauer, K.; Von Zahn, M.; Hinz, O. Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Inf. Syst. Res.* **2023**, *34*, 1582–1602. [CrossRef]
109. Chatterjee, S.; Shenoy, P. Model-agnostic fits for understanding information seeking patterns in humans. *arXiv* **2020**. [CrossRef]
110. Gajos, K.Z.; Mamykina, L. Do people engage cognitively with AI? Impact of AI assistance on Incidental Learning. In Proceedings of the 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, 22–25 March 2022; pp. 794–806. [CrossRef]
111. Huang, Y.; Cheng, Y.; Chen, L.; Hsu, J.Y.-J. Human-AI Co-Learning for Data-Driven AI. *arXiv* **2019**. [CrossRef]

112. Russell, S.; Moskowitz, I.S.; Raglin, A. Human information interaction, artificial intelligence, and errors. In *Autonomy and Artificial Intelligence: A Threat or Savior?* Springer: Berlin/Heidelberg, Germany, 2017; pp. 71–101. [CrossRef]
113. Fisher, M.; Smiley, A.H.; Grillo, T.L.H. Information without knowledge: The effects of Internet search on learning. *Memory* **2021**, *30*, 375–387. [CrossRef]