Research Ethics: Journal of Multidisciplinary Research
ISSN: 2584-2552
Vol. 2, Issue 2, March-April 2024
www.researchethics.in

Addressing Ethical Dilemmas in Conversational Interfaces: A Review of ChatGPT's Implications

Kaajal Sharma

Research Scholar, University Business School, Panjab University, Chandigarh, India.
Email: kaajalsharma55@gmail.com

Abstract

Conversational interfaces powered by AI, like ChatGPT, have experienced a remarkable surge in popularity, revolutionizing the way users interact with technology. These interfaces are capable of facilitating conversations that closely resemble human-like interactions, thus enhancing user experiences across various applications. However, the widespread deployment of such AI-powered conversational systems also introduces a host of ethical considerations that cannot be overlooked. As these technologies become increasingly prevalent in our daily lives, it becomes imperative to address these ethical concerns to ensure that their impact is positive and aligns with societal values. The responsible and ethical use of conversational AI systems demands careful attention to issues such as bias and fairness, misinformation, user privacy, transparency, psychological impact, accountability, and responsibility. By proactively addressing these ethical considerations, we can harness the potential of conversational interfaces while ensuring that they uphold ethical principles and promote a user-centric and inclusive technological landscape. This review paper examines the ethical implications of ChatGPT for conversational interfaces, addressing bias, fairness, misinformation, user privacy, transparency, psychological impact, and accountability. It emphasizes the need to mitigate biases, fact-check information, handle user data responsibly, obtain user consent, prioritize user well-being, and engage in an iterative improvement process. The review concludes by advocating for an iterative improvement process that incorporates user feedback and engages a multidisciplinary approach involving AI researchers, ethicists, psychologists, and end-users.

Keywords: ChatGPT, conversational interfaces, ethics, user privacy, data security, accountability.

1. Introduction

In recent years, the widespread adoption of conversational interfaces powered by AI, such as ChatGPT, has revolutionized the way we interact with technology. These interfaces have gained immense popularity due to their ability to simulate human-like conversations, providing users with a more natural and engaging experience. By bridging the gap between humans and machines, AI-powered conversational systems have transformed various domains, including customer service, virtual assistants, and even entertainment.

However, alongside their remarkable advancements, the rise of conversational AI systems has raised significant ethical concerns that demand attention. As these technologies become increasingly integrated into our daily lives, it is imperative to address the ethical implications they present to ensure that their impact aligns with societal values and promotes positive outcomes.

One of the primary ethical considerations surrounding conversational interfaces is the presence of bias and the need for fairness. AI models, including ChatGPT, can inadvertently perpetuate biases present in training data, leading to discriminatory or unfair responses. Recognizing and mitigating these biases becomes crucial to ensure impartiality and equal treatment in conversations.

Another critical concern is the spread of misinformation through AI-powered systems. As conversational interfaces gain influence and authority, they have the potential to disseminate inaccurate or misleading information. To maintain integrity and trust, it is vital to implement mechanisms that verify information sources, fact-check responses, and promote accurate and reliable information exchange.

The issue of user privacy arises as conversational AI systems collect and process personal data during interactions. Safeguarding user privacy and ensuring responsible handling of sensitive information is paramount to establish user trust and protect individuals from potential data misuse or breaches.

Transparency in AI systems is another key aspect that must be addressed. Users should have a clear understanding of when they are interacting with an AI system, as opposed to a human, and be informed about the system's limitations and capabilities. Transparent communication builds trust and empowers users to make informed decisions.

Furthermore, the psychological impact of conversational AI on users cannot be overlooked. Human-like interactions may lead users to form emotional connections or rely on AI systems for emotional support. Ensuring user well-being, mental health, and preventing any potential harm or dependency are crucial considerations in the ethical development and deployment of conversational interfaces.

Accountability and responsibility are integral to the ethical use of AI systems. Developers and operators must be held accountable for the outcomes and impact of their systems, taking responsibility for addressing any unintended consequences or harmful effects that may arise.

To navigate these complex ethical challenges, this review paper proposes an iterative improvement process that involves a multidisciplinary approach. Collaboration between AI researchers, ethicists, psychologists, and end-users is crucial in addressing these concerns effectively. Additionally, incorporating user feedback throughout the development and deployment lifecycle is vital in continuously improving the ethical performance and user experience of conversational AI systems.

By proactively addressing the ethical considerations associated with ChatGPT and similar conversational AI systems, we can ensure that these technologies are developed and deployed in a manner that upholds ethical principles and aligns with societal values. This comprehensive review serves as a call to action, advocating for a collective effort to navigate the ethical challenges and create a user-centric and inclusive technological landscape.

2. Literature Review

In this comprehensive literature review, the focus is on ethical dilemmas surrounding conversational interfaces such as ChatGPT and other AI tools. A thorough examination of thirty relevant studies has been conducted, with careful selection from esteemed authors, reputable publishers, renowned journals, and reputable conference proceedings. To provide a clear framework for analysis, the research has been organized into two distinct phases: the period before the release of ChatGPT and the period after its release. This categorization allows for a comprehensive understanding of the evolution of ethical concerns in the field of conversational interfaces, highlighting the impact and implications of ChatGPT on the discourse surrounding ethical considerations.

2.1 Period Before the Release of ChatGPT

The rise of ChatGPT has sparked renewed attention to the ethical considerations surrounding its use. However, it's important to recognize that ethical concerns in technology have been studied for decades. In the 2000s, Friedman et.al., (2002) suggested the concept of Value Sensitive Design, which focuses on incorporating human values into the technology design process. This approach involves conceptual, empirical, and technical investigations to identify stakeholders, understand values, study the human context, and assess the impact of design. Examples from various domains highlighted the significance of clarifying values during design. As technological revolutions occur, unique ethical challenges arise.

Moor (2005) suggested that policies and ethical frameworks need to be developed to address these issues. Examples such as wireless computing and genetic, nanotechnology, and neurotechnology advancements demonstrate the complexities involved in formulating and justifying new policies. The convergence of technologies increases the importance of ethics in navigating the growing number of ethical problems.

Briggle (2009) conducted a study aimed to define computing ethics thematically, highlighting interdisciplinary collaboration, methodical study, practical impact, and global society as key themes. The findings can benefit computing ethicists, educators, and establish a theoretical framework for integrating ethics into computing.

Mittelstadt et al., (2016) proposed a conceptual map of the ethics of algorithms, identifying concerns related to evidence, outcomes, effects, and traceability. The map offers a framework for discussing ethical issues but doesn't provide solutions. Specific concerns include reliance on uncertain knowledge, opaque processes, biased outcomes, and transformative effects on society. Traceability is emphasized for identifying causes and responsibilities in ethical failures.

Caliskan et.al., (2017) conducted a study demonstrating that word embedding can replicate implicit biases measured by the Implicit Association Test (IAT). Machine learning can absorb biases without direct experience or explicit representation of semantics, raising concerns about the perpetuation of biases.

Cadwalladr and Graham, (2018) in the "Social Challenges of AI" focus on AI integration into social institutions, inequality, political populism, and industry scandals and discuss the Gaps in AI ethics and accountability, along with scandals involving tech companies like Facebook data breach.

Conger and Cameron, (2018) explored Google's involvement in the Department of Defense's drone surveillance program, which sparked significant dissent within the technology industry. Wakabayashi, (2018) examined the consequences of AI systems tested on live populations, including autonomous car accidents and erroneous visa cancellations. The study emphasized the need for greater accountability, public oversight, and due process in addressing these challenges.

Ram (2018) highlighted the barriers to accountability posed by industrial and legal secrecy in AI development and the incentives driving rapid technical AI research. The study concluded with substantive approaches and recommendations for addressing these issues and fostering greater accountability in AI systems within a wider social context. Till now the need for greater accountability, public oversight, and due process is emphasized, along with the challenges posed by secrecy in AI development.

A simple solution to these challenges was proposed in 2019. Mitchell et al., (2019) proposed the use of model cards as a framework for transparent model reporting. These cards provide benchmarked evaluation and relevant information about machine learning models. The adoption of model cards promotes transparency and greater accountability. Challenges faced by the fair-ML community in developing machine learning systems that achieve fairness and justice are discussed. The study highlighted the importance of considering the societal context and human involvement while developing machine learning systems.

Weber (2020) addressed several questions related to AI processes, including compliance with human rights and non-discrimination, the legal basis for automated decision-making, adherence to data protection laws, and responsibility for monitoring and liability. He emphasized the importance of transparency, accountability, and safety in AI systems and advocated for regulatory tools to minimize technological risks and place humans at the center of AI deployment. Also, the author suggested that transparency should focus on disclosing the logic of algorithms, while accountability entails responsibility and justification for actions. Safety and robustness are identified as crucial for trust in AI systems. It was concluded that there is a need to balance regulation to safeguard public trust, fundamental rights, personal self-determination, and non-discrimination in AI development and deployment.

Casebeer (2020) distinguishes between the ethics of AI, which deals with applying AI to specific domains, and AI ethics, which focuses on using AI to develop ethical systems. The paper delves into the concept of an artificial conscience and its role in decision-making, emphasizing the importance of moral sensitivity, judgment, motivation, and skill. The author argues that building an artificial conscience is not only ethically permissible but also morally obligated due to the significant power wielded by autonomous algorithms and platforms. The paper highlights four reasons for developing an artificial conscience: evolutionary trends in the use of autonomy in warfare, military doctrine requirements, national security concerns, and moral obligations. The author concludes that an artificial conscience can contribute to more ethical and effective decision-making in autonomous systems. This paper provides valuable insights into the intersection of AI and ethics, shedding light on the implications and considerations for developing ethical AI systems. The discussion on the ongoing debate surrounding the ethics of artificial intelligence and other advanced technologies that heavily rely on data was also highlighted at that time.

Raab (2020) acknowledges the importance of ethics in technology development and the need for a impact assessment that goes beyond traditional privacy concerns. However, the author also highlighted the challenges posed by the multitude of ethical frameworks and their lack of clarity and stability and emphasized the importance of cultivating judgment and incorporating social and organizational governance mechanisms to ensure valid and trustworthy assessments. It was suggested that impact assessments should be viewed as an element of technology governance and call for regulatory approaches that can adapt to rapid change and remain uniform across different innovations. The author also proposed integrating impact assessment into institutional practices and research approval systems to enhance ethical values and support technological development.

Alujevic et al., (2020) emphasized the need for integrated approaches involving governments, industry, and academia, as well as multidisciplinary frameworks and public engagement to address the complex technical aspects of the debate. The authors concluded that while policy documents generally align with their vision of a more ethical AI, they lack comprehensive and interconnected approaches, focusing more on ethical frameworks than on regulatory possibilities. They highlighted the need for new regulatory frameworks, the codification of privacy requirements, the establishment of standards, and the promotion of multi-stakeholder forums and public deliberation to ensure the responsible development and use of AI.

Mantelero and Esposito (2021) demonstrated through analysis of decisions and documents from data protection authorities that human rights already play a role in regulating data use. They proposed a Human Rights Impact Assessment (HRIA) methodology and model specifically tailored for AI applications, which allows for a more measurable approach to risk assessment. The authors suggested that the proposed model can guide AI developers, municipalities, governments, and private companies in the development and deployment of AI products and services while ensuring human rights are respected. The model can also be used by supervisory authorities and auditing bodies to monitor the impact of data use on individual rights and freedoms. The authors argued that conducting HRIA should be seen as an opportunity rather than a burden and that it can facilitate the development of human-centric AI and standardized assessment of AI solutions.

Bender et.al., (2021), explored the potential of algorithmic decision-making enabled by AI and machine learning to improve societal decision-making. Risks such as privacy violations, power imbalances, lack of transparency and accountability, and discrimination and bias are identified. Multidisciplinary collaboration is deemed crucial to effectively address these limitations and ensure the ethical development and deployment of AI systems. An example of such multidisciplinary collaboration is ChatGPT.

2.2 Period After the Release of ChatGPT

With the introduction of ChatGPT, the focus of studies shifted towards the ethical validity of ChatGPT as a conversational interface. Karaarslan (2022) collected abstracts of relevant papers from 2020 to 2022 and paraphrased them using ChatGPT. They also asked ChatGPT-specific questions related to the topic. The results showed promising potential for AI assistance in the literature review process, although there were significant matches detected when comparing the paraphrased parts with the Ithenticate plagiarism tool. The study highlights the acceleration of knowledge compilation and expression with the help of AI and suggests that future academic publishing processes could require less human effort. Further analysis and monitoring of citations to evaluate the academic validity of the content generated by ChatGPT are planned for future studies. The findings demonstrate both the capabilities and limitations of using AI in literature review tasks.

Varona and Suarez (2022) analyzed the concepts of "Discrimination," "Bias," "Fairness," and "Trustworthiness" as variables in the social impact of artificial intelligence (AI). They examined how these variables relate to the principles outlined in the Principled AI International Framework and highlighted the lack of consensus within the scientific community regarding their standardization. The study emphasized the interdependency between bias and discrimination and discussed their implications in algorithmic decision-making systems (ADMS). It identifies the role of data gathering, cleaning, processing, and the development team's biases in contributing to biased and discriminatory outcomes. The authors proposed that trustworthiness in ADMS should be built upon fairness and non-discrimination, and they suggested four derived features, including transparency, security, project governance, and bias management, as checkpoints for achieving capability and maturity in developing trustworthy AI systems. The study aimed to provide a methodological reference tool, specifically a Capability and Maturity Model, to support software engineers, particularly ADMS developers, in incorporating social and ethical dimensions into their work. Meanwhile, a new threat was found.

Goel (2022) discussed ethical concerns related to artificial intelligence (AI). He highlighted three broad categories of concerns: the fear of super-intelligent machines surpassing human intelligence and potentially harming humans, biases in data and algorithms that can lead to unfair outcomes, and the intentional use of AI for malicious purposes. The author argued that human involvement is central to all these concerns, emphasizing the need for a social and cultural perspective on intelligence. Additionally, he introduced a fourth category of concern, the abuse of AI by humans, highlighting the potential for mistreatment and exploitation of AI agents. The author suggested that understanding and addressing these ethical concerns require interdisciplinary collaboration and the development of responsible AI systems.

Neiderman and Baker (2022) reflected on the unique ethical issues that arise from the use of artificial intelligence (AI) in information systems (IS) applications. It categorized AI ethics issues into three distinct categories: viewing AI as an IS application, as a generative capacity producing unpredictable outputs, and as a basis for reexamining the nature of mental phenomena. The authors explored these categories and discussed the potential emergence of ethical issues as AI capabilities advance. It also delves into the relationship between consciousness and agency in AI, discussing the possibility of machine-generated consciousness and the implications it may have on human-AI interactions. The authors emphasized the importance of anticipating and addressing ethical issues in AI as technology progresses. They analyzed research patterns, citation relationships among researchers, and highly referenced journals in the field. The findings reveal that the United States is the leading contributor to AI and ethics research, followed by Western Europe and East Asia. The top ten nations accounted for the majority of publications, with the United States producing the highest number of papers. Switzerland exhibited the highest research production adjusted for population size.

Chuang et.al., (2022) highlighted the evolving nature of AI and ethics research over the past 70 years and emphasizes the dominance of developed countries in this field. They also pointed out the prevalence of ethical issues in engineering-related AI applications. The authors concluded by suggesting that understanding the development and trends in AI and ethics research is crucial for anticipating future implications and fostering meaningful discussions in the field.

Awad et.al., (2022) introduced the concept of computational ethics as a framework to address the ethical challenges posed by AI and machine learning. It emphasized the need to incorporate the study of human moral decision-making into the development of ethical AI systems. The framework aimed to inform the engineering of AI systems and understand human moral judgment in computational terms. By integrating diverse research questions and collaborating across multiple academic communities, computational ethics can shed light on longstanding philosophical questions and facilitate the development of ethical AI systems.

Canas (2022) emphasized the importance of considering the collaboration between human beings and AI systems as a shared responsibility. The concept of co-supervision is introduced, highlighting that each agent should supervise and be aware of the actions of the other to ensure the achievement of activity objectives and share responsibility for the consequences. The author suggested that understanding the boundaries and factors influencing human supervision of AI actions, as well as utilizing psychological research on external observer supervision of human actions, can help establish shared responsibility and ethical collaboration. He also suggested that incorporating the concepts of accountability and responsibility in the context of collaboration between humans and AI systems is essential for designing ethical AI systems and achieving the objectives of the ongoing social debate on AI and ethics.

With the frequent usage of ChatGPT for various purposes, new challenges arrived. Alser and Waisberg (2023) expressed concerns regarding the increasing use of ChatGPT, an artificial intelligence language model, in academia and medicine. The authors highlighted the issue of ChatGPT being credited as an author in medical articles, which goes against the guidelines set by the International Committee of Medical Journal Editors (ICMJE) for authorship eligibility. They discussed how ChatGPT's contributions to the writing of these papers do not fulfill the criteria for authorship and raised questions about the significance of the chatbot's approval in the publication process. The authors also addressed the problem of plagiarism, noting that ChatGPT has been found to copy content from unreliable sources without proper citation. Additionally, they discussed biases in ChatGPT's output, the lack of transparency regarding its learning data sources, and the potential for manipulation of its responses by developers and users. The authors cautioned against using ChatGPT in academia and scientific publications due to its limitations in-depth, factual accuracy, and ethical concerns related to plagiarism and biases.

Salvagno et.al., (2023) discussed the use of ChatGPT, a chatbot generative pre-trained transformer developed by OpenAI, in scientific writing. The authors highlighted that ChatGPT can assist in various tasks such as drafting articles, summarizing data, and providing language reviews. It has the potential to make scientific writing faster and easier, particularly in tasks like automated draft generation, article summarizing, and language translation. However, the use of AI chatbots in scientific writing raises ethical concerns and should be regulated. The authors emphasized that while chatbots can be helpful, they should not replace human researchers' expertise, judgment, and responsibility. They also explored the limitations of chatbot applications in scientific writing and discussed potential ethical considerations. The author also emphasized the need for international academic regulations regarding the use of chatbot tools in scientific writing.

Iskender (2023) explored the impacts of OpenAI's ChatGPT on higher education and academic publishing. The unique aspect of his study is that ChatGPT itself is interviewed as the subject. The results of the interview revealed that ChatGPT can be used to delegate monotonous tasks such as grading, allowing instructors to focus on more intellectual activities. Students can also utilize ChatGPT for brainstorming ideas. However, the author acknowledged the risks of over-reliance on ChatGPT, which may diminish critical thinking skills and contribute to educational inequalities. ChatGPT also stated that it cannot replace human creativity and originality in academic work. Additionally, the author suggested that ChatGPT can be beneficial in the tourism and hospitality industry for personalized services and content creation.

Dwivedi et.al., (2023) explored the transformative potential of generative AI tools like ChatGPT, which can generate text that resembles human-produced content. It discusses the wide range of applications and the opportunities and challenges associated with these tools. The authors highlighted the capabilities of ChatGPT to enhance productivity and benefit industries like banking, hospitality, tourism, and information technology. However, they also addressed limitations, including disruptions to existing practices, threats to privacy and security, and the consequences of biases, misuse, and misinformation. They emphasized the importance of responsible use and identified areas for further research, such as knowledge, transparency, ethics, digital transformation, and education. The implications for practice and policy are discussed, including the need for organizational changes, criteria to evaluate outputs, combating resistance to change, addressing limitations, and developing regulations and guidelines to govern the use of generative AI tools like ChatGPT. The authors also highlighted the challenges of biases in AI systems and the responsibility of AI practitioners and users to mitigate them. They concluded by emphasizing the significant opportunities and challenges presented by ChatGPT and the need for laws and international coordination to maximize its benefits while addressing ethical and practical concerns.

Ray (2023) focused on its background, applications, challenges, and prospects. The author discussed the origins and development of ChatGPT, and its various applications in industries such as customer service, healthcare, and education, and highlights critical challenges faced by the model. These challenges include reliability and accuracy, bias in AI models, overreliance on AI, quality control, dataset bias, generalization, explainability, energy consumption, real-time responsiveness, safety concerns, privacy concerns, cultural and linguistic bias, model explainability, adapting to domain-specific knowledge, contextual understanding, and factual accuracy. The author also addressed ethical considerations surrounding ChatGPT, including data privacy and security, transparency and accountability, bias and fairness, misuse and abuse, responsibility and accountability, adversarial attacks, misinformation, autonomy, human-like interactions, environmental impact, and bias and discrimination. The author stressed the importance of proactive approaches to address these challenges and ethical concerns, ensuring the responsible development and use of AI language models like ChatGPT. Despite the controversies and ethical concerns, the author recognized ChatGPT's remarkable potential for revolutionizing scientific research and envisions a future where it is integrated with other technologies, improves human-AI interaction, and addresses the digital divide.

Rahman et. al., (2023) conducted a practical example using a research topic and assessed the capabilities and limitations of ChatGPT in writing an academic paper. They found that ChatGPT can be effective for idea generation, outlining research topics, and writing abstracts using prompts. It can also summarize large amounts of text and identify key findings from the literature. However, the authors observed limitations in writing sections such as the research problem, literature review, and statistical analysis. ChatGPT generated hypothetical statements, fake citations, and lacked access to real datasets. The authors recommended using ChatGPT as a complementary tool, not as the sole means of writing a research article and emphasized the need for human control and accountability. While ChatGPT can improve research efficiency, researchers should be cautious and verify the accuracy and reliability of the information it provides. The findings of the authors have important implications for the responsible use of ChatGPT and highlight the need for guidelines and transparency in its application in academic research.

Sok and Heng (2023) discussed the benefits and risks associated with using the Generative Pre-trained Transformer (ChatGPT) AI tool in education and research. The authors highlighted five main benefits of ChatGPT, including creating learning assessments, enhancing pedagogical practices, offering virtual personal tutoring, generating academic outlines, and brainstorming ideas. However, they also acknowledged risks related to academic integrity, unfair learning assessment, inaccurate information, and over-reliance on AI. The authors have provided recommendations for the effective use of ChatGPT, such as promoting inclusive and ethical use, revising assessment standards, providing training for teachers and students, conducting action research, and being vigilant in verifying the accuracy of generated responses. It advises researchers to utilize ChatGPT for a better understanding of its advantages and flaws but cautions against using it to produce entire research articles to prevent academic misconduct.

Khogali and Mekid (2023) provided a comprehensive analysis of the potential impacts of artificial intelligence (AI) and machine learning on society. They discussed the positive implications and drawbacks of AI technology in various industries such as transportation, health, education, and the environment. The authors investigated the long-term consequences of AI on human civilization, including concerns such as fear of AI, job losses, dehumanization of jobs, and the impact on employees' well-being and highlighted the misconceptions and fears associated with AI, as well as the potential risks such as unemployment and societal inequality. The authors emphasized the need for education, training, and careful implementation of AI to ensure a qualified workforce and prevent negative consequences. Additionally, they explored the safety and acceptance concerns related to autonomous vehicles.

Table 1: Studies related to factors affecting ethical considerations.

Factors Affecting Ethical Considerations	Related Studies
Value sensitive Design	Friedman et.al., (2002)
Policy and Ethical Framework	Moor (2005), Alujevic et.al. ,(2020), Raab (2020), Dwivedi et.al., (2023)
Ethical Issues with Algorithms	Raab (2020), Bender et.al., (2021)
Implicit Biases in AI Systems	Mittelstadt et al., (2016), Caliskan et.al., (2017), Mantelero and Esposito (2021), Varona and Suarez (2022), Goel (2022), Alser and Waisberg (2023), Ray (2023)
Accountability	Cadwalladr and Graham, (2018), Wakabayashi, (2018), Ram (2018), Mitchell et al., (2019), Weber (2019), Canas (2022)
Transparency	Mitchell et al., (2019), Weber (2019), Mantelero and Esposito (2021), Varona and Suarez (2022)

Standardization	Alujevic et.al.,(2020), Varona and Suarez (2022), Sok and Heng (2023)
Safety	Weber (2019), Ray (2023), Khogali and Mekid (2023)
Robustness	Weber (2019)
Human Rights and Non-Discrimination	Weber (2019), Mantelero and Esposito (2021)
Collaboration	Briggle (2009), Mantelero and Esposito (2021), Goel (2022), Awad et.al., (2022)
Multi-Stakeholder Engagement	Alujevic et.al., (2020), Mantelero and Esposito (2021)
Regulation and Governance	Weber (2019), Raab (2020)
Integration of Human Moral Decision-Making	Friedman et.al., (2002), Mitchell et al., (2019), Mantelero and Esposito (2021), Goel (2022), Awad et.al., (2022)
Shared Responsibility	Neiderman and Baker (2022), Awad et.al., (2022), Ray (2023)
Academic Validity and Plagiarism	Karaarslan (2022), Awad et.al (2022), Alser and Waisberg (2023), Rahman et.at., (2023), Sok and Heng (2023)
International Academic Regulations	Salvagno et.al., (2023)
Educational Implications	Iskender (2023)
Fairness and Justice	Mitchell et al., (2019), Varona and Suarez (2022), Ray (2023)
Compliance and Data protection	Mitchell et al., (2019), Mantelero and Esposito (2021)

- **Source** - Authors' compilation based on Literature Review

3. Findings

The review conducted in this study has identified a comprehensive set of twenty factors that impact ethical considerations, as outlined in Table 1. Within the body of literature under review, it becomes evident that eight factors have consistently garnered substantial citations. These factors hold significant importance and include the following: integration of human morale decision-making, policy and ethical framework, transparency, accountability, implicit bias, collaboration, standardization, and academic validity and plagiarism. Repeated emphasis on these factors highlights their critical role in shaping ethical considerations within the context of the studies examined.

3.1 Integration of Human Moral Decision-Making

The incorporation of human values and ethical considerations into the design and development of AI systems is very crucial. This requires involving stakeholders, understanding their values, and assessing the impact of AI on human decision-making. It is evident from the research that incorporating human values in the technology design process leads to better human-technology connection and shared decision-making (Freidman et.al., 2002). Human involvement in the development of machine learning systems leads to lesser societal impacts. (Mitchell et.al.,2019). Also, the use of Human Right Impact assessment while developing AI technology leads to the development of more human-centric AI tools (Mantelero and Esposito, 2021).

3.2 Policy and Ethical Framework

The development and implementation of policies and ethical frameworks that address the unique challenges posed by AI technologies is the need of the hour. This includes formulating and justifying new policies to navigate the complexities of emerging technologies. The convergence of technologies increases the importance of ethics to navigate the growing number of ethical problems and the requirement of valid changes in policies and frameworks to adapt to the changes (Moor, 2005). Specific concerns include reliance on uncertain knowledge, opaque processes, biased outcomes, and transformative effects on society. For this purpose, a framework is required to address the problems (Mittelstadt et al., 2016).

3.3 Transparency

It is crucial to promote transparency in AI systems. This involves disclosing the logic of algorithms, ensuring accountability for actions, and providing benchmarked evaluation and relevant information about machine learning models through tools like model cards. The adoption of model cards promotes transparency and greater accountability (Mitchell et. al., 2019). It was proposed that transparency should focus on disclosing the logic of algorithms (Weber, 2019). Algorithmic decision-making systems (ADMS) should be built upon the suggested four derived features, including transparency, security, project governance, and bias management, as checkpoints for achieving capability and maturity in developing trustworthy AI systems (Varona and Suarez, 2022).

3.4 Accountability

There should be an emphasis on the need for greater accountability in AI systems. This includes public oversight, due process, and the identification of causes and responsibilities in ethical failures. Regulatory tools can be implemented to minimize technological risks and ensure accountability. It was suggested that incorporating the concepts of accountability and responsibility in the context of collaboration between humans and AI systems is essential for designing ethical AI systems and achieving the objectives of the ongoing social debate on AI and ethics (Canas, 2022). It was also suggested that transparency should focus on disclosing the logic of algorithms, while accountability entails responsibility and justification for actions (Weber, 2019). A researcher also emphasized the need for greater accountability, public oversight, and due process in addressing these challenges (Wakabayashi, 2018). The barriers to accountability posed by industrial and legal secrecy in AI development and the incentives driving rapid technical AI research are also highlighted (Ram, 2018).

3.5 Implicit Bias

It is needed to be aware of the potential for biases in AI systems. Machine learning models can replicate implicit biases, leading to biased outcomes. It is important to address these biases through data collection, processing, and bias management to ensure fairness and non-discrimination in algorithmic decision-making (Caliskan et.al., 2017). The researchers explored the potential of algorithmic decision-making enabled by AI and machine learning to improve societal decision-making and reduce the implicit bias (Bender et.al., 2021). The researchers also emphasized the interdependency between bias and discrimination and discussed their implications in algorithmic decision-making systems (ADMS). It identifies the role of data gathering, cleaning, processing, and the development team's biases in contributing to biased and discriminatory outcomes (Varona and Suarez 2022).

3.6 Collaboration

There should be a focus on interdisciplinary collaboration to address ethical challenges in AI development and deployment. This involves collaboration between academia, industry, and government, as well as public engagement. Multi-stakeholder forums and deliberation can contribute to the responsible development and use of AI (Canas, 2022). The authors also suggested that understanding and addressing these ethical concerns require interdisciplinary collaboration and the development of responsible AI systems (Goel, 2022).

3.7 Standardization

It should be taken care that the standardization of ethical frameworks and practices in AI development is considered. This includes the establishment of standards, the codification of privacy requirements, and the development of capability and maturity models to ensure trustworthy and ethical AI systems (Varona and Suarez, 2022). The researchers also emphasized the need for international academic regulations regarding the use of chatbot tools in scientific writing (Salvagno et.al., 2023). It was proposed that there is a need to balance regulation to safeguard public trust, fundamental rights, personal self-determination, and non-discrimination in AI development and deployment (Weber, 2019).

3.8 Academic Validity and Plagiarism

It is needed in academia and scientific publishing to carefully evaluate the use of AI language models like ChatGPT. Concerns regarding authorship, plagiarism, biases, and transparency should be addressed. Academic regulations and guidelines should be established to ensure the responsible and ethical use of AI tools in scientific writing (Karaarslan, 2022). The authors cautioned against using ChatGPT in academia and scientific publications due to its limitations in-depth, factual accuracy, and ethical concerns related to plagiarism and biases (Alser and Waisberg, 2023). The researchers also highlighted the acceleration of knowledge compilation and expression with the help of AI and suggested that future academic publishing processes could require less human effort. Further analysis and monitoring of citations to evaluate the academic validity of the content generated by ChatGPT are very important before writing (Bender et.al., 2021).

Conclusion

In conclusion, this study highlights several key findings related to ethical considerations in AI development and deployment. Incorporating human moral decision-making, policy and ethical frameworks, transparency, accountability, addressing implicit bias, fostering collaboration, standardization, and ensuring academic validity and plagiarism concerns are crucial factors to prioritize. The integration of human values into AI design leads to better human-technology connections and shared decision-making. The establishment of policies and frameworks helps navigate the complexities and ethical challenges posed by AI technologies. Promoting transparency through disclosing algorithm logic and ensuring accountability is essential. Addressing implicit bias and fostering interdisciplinary collaboration can contribute to the responsible development and use of AI. Standardization of ethical frameworks and addressing academic validity and plagiarism concerns are necessary. Ultimately, users should prioritize ethical considerations, be cautious of limitations and risks, and use AI tools appropriately in academic and scientific contexts to ensure responsible and ethical AI development and deployment.

References

Alser, M., & Waisberg, E. (2023). Concerns with the usage of ChatGPT in Academia and Medicine: A viewpoint. *Am. J. Med. Open, 100036*. doi: https://doi.org/10.1016/j.ajmo.2023.100036

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., & Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences*.

Bender, Emily, et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Timnit Gebru Timnit@Blackinai.org Black in AI Palo Alto, CA, USA CCS

CONCEPTS • Computing Methodologies → Natural Language Processing. ACM Reference Format." *FAccT '21, March 3–10, 2021, Virtual Event, Canada*, 1 Mar. 2021, faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf, https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf

Briggle, A. (2009). Computer-mediated friendship: Illustrating three tasks for a computer ethics of the good. In *Proceedings of the 8th International Conference CEPE* (pp. 135-147).

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 Million Facebook

Profiles Harvested for Cambridge Analytica in Major Data Breach. The Guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Cañas, J. J. (2022). AI and Ethics when human beings collaborate with AI Agents. *Frontiers in psychology*, *13:836650*. doi: 10.3389/fpsyg.2022.836650

Casebeer, W. D. (2020). Building an Artificial Conscience: Prospects for Morally Autonomous Artificial Intelligence. In *Artificial Intelligence and Global Security* (pp. 81-94). Emerald Publishing Limited.

Chuang, C. W., Chang, A., Chen, M., Selvamani, M. J. P., & Shia, B. C. (2022). A Worldwide Bibliometric Analysis of Publications on Artificial Intelligence and Ethics in the Past Seven Decades. *Sustainability*, *14*(18), 11125.

Conger, K., & Cameron, D. (2018, March 6). Google Is Helping the Pentagon Build AI for Drones. Gizmodo. https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Friedman, B., Kahn Jr, P. H., & Borning, A. (2002). Value Sensitive Design and Information Systems. Human–Computer Interaction, 16(4), 247-268.

Goel, A. (2022). Looking back, looking ahead: Humans, ethics, and AI. *AI Magazine*, *43*(2), 267-269.

Iskender, A. (2023). Holy or unholy? Interview with open AI's ChatGPT. *European Journal of Tourism Research*, *34*, 3414-3414.

Khogali, H. O., & Mekid, S. (2023). The blended future of automation and AI: Examining some long-term societal and ethical impact features. *Technology in Society*, *73*, 102232.

Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review*, *41*, 105561.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., & Gebru, T.

(2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

Moor, J. H. (2005). Why we need better ethics for emerging technologies. *Ethics and information technology*, *7*(3), 111-119.

Niederman, F., & Baker, E. W. (2023). Ethics and AI Issues: Old Container with New Wine?. *Information Systems Frontiers*, *25*(1), 9-28.

Raab, C. D. (2020). Information privacy, impact assessment, and the place of ethics. *Computer Law & Security Review*, *37*, 105404.

Ram, N. (2018). Innovating Criminal Justice. Northwestern University Law Review, 112(4), 659–724. https://scholarlycommons.law.northwestern.edu/nulr/vol112/iss4/2

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Russell, S. J., & Norvig, P. (1995). Artificial Intelligence: A Modern Approach. Prentice Hall.

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing?. *Critical care*, *27*(1), 1-5.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68).

Varona, D., & Suárez, J. L. (2022). Discrimination, Bias, Fairness, and Trustworthy AI. *Applied Sciences*, *12*(12), 5826. https://doi.org/10.3390/ app12125826

Vesnic-Alujevic, L., Nascimento, S., & Polvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy*, *44*(6), 101961.

Wakabayashi, D. (2018, July 30). Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. *The New York Times*. https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

Weber, R. H. (2020). Socio-ethical values and legal rules on automated platforms: The quest for a symbiotic relationship. *Computer Law & Security Review*, *36*, 105380.