Annals of Engineering Mathematics and Computational Intelligence (AEMCI)
Vol. 1, 2024, pp. 1~12

Exploring Ethical Implications of ChatGPT and Other AI Chatbots and Regulation of Disinformation Propagation

Glorin Sebastian[1], Dr. Shaliet Rose Sebastian [2]
[1]Georgia Institute of Technology, Atlanta, GA 30332
[2]Associate Professor in Community Medicin Kerala University of Health Sciences, India
gsebastian6@gatech.edu, glorin17xime@gmail.com

Article Info

Article history:

Received July 08, 2024
Revised August 15, 2024
Accepted September 10, 2024

Keywords:

Ethical AI
Algorithmic Fairness
ChatGPT
Disinformation
AI Governance
Deep Fakes

ABSTRACT

Chatbots, driven by artificial intelligence (AI) such as ChatGPT, are playing an increasingly pivotal role in the digital age and are disrupting numerous industries. As these technologies rapidly advance and their influence expands, they present a range of ethical and regulatory challenges. One critical concern is the potential use of these AI chatbots to disseminate disinformation. Given the capacity of these systems to generate text that closely mimics human conversation based on training data, there is a significant risk of them being manipulated to widely disseminate inaccurate or deceptive information. Such misuse could result in various societal problems, including the intensification of political divisiveness or the spreading of damaging misinformation. This paper embarks on a critical exploration of these ethical issues, with a specific focus on the potential misuse of AI chatbots in the propagation of disinformation. This research further investigates potential regulatory interventions that could alleviate these issues. In the rapidly evolving world of AI technology, creating a robust regulatory framework that balances the benefits of AI chatbots with the prevention of their misuse is crucial. Therefore, this paper aims to contribute to the ongoing dialogue about the ethical use of AI and the development of effective regulatory strategies.

Corresponding Author:

Glorin Sebastian
Georgia Institute of Technology
Atlanta,
GA 30332
Email: gsebastian6@gatech.edu

1.0 Introduction

Over the past decade, there has been a tremendous increase in the use of artificial intelligence (AI) chatbots. Their extensive influences change conventional human-computer interactions and penetrate many areas of our daily life, such as customer service, education, and mental health support. These AI systems' deep advantages and benefits are beyond dispute, and they continue to offer enormous promise for further social breakthroughs. Such inaccurate or deceptive information can have serious repercussions, such as swaying public opinion or igniting social unrest, worsening the information integrity dilemma in our increasingly digital society. Weidinger et al. systematically structured the ethical risk landscape with LLMs, clearly identifying six risk areas: 1) Discrimination, Exclusion, and Toxicity, 2) Information Hazards, 3) Misinformation Harms, 4) Malicious Uses, 5) Human-Computer Interaction Harms, 6) Automation, Access, and Environmental Harms.

There have been numerous proposals for thorough regulatory measures and a closer ethical investigation in response to these ethical concerns about the role of AI chatbots in disinformation dissemination. Establishing a strong legislative framework that adequately handles these problems is of utmost relevance in a world where AI is developing quickly. It is also vital to have thoughtful ethical debates and evaluations of these technologies.

This study intends to explore the ethical implications of AI chatbots, particularly ChatGPT and others of its sort, in light of these difficulties. In order to contribute to the larger discussion on the responsible and ethical use of AI technology, it will continue to examine various regulatory remedies to the issue of disinformation propagation.

2.0 Ethical Implications

2.1 Disinformation Propagation:

AI chatbots, due to their design and capacity, may propagate disinformation including deepfakes, which can have serious consequences for individuals and society (Derner, E., & Batistič, K., 2023). Disinformation, a frequently employed tactic aimed at manipulating public opinion, has the potential to undermine trust, escalate social divisions, and even incite acts of violence (Lewandowsky, Ecker, & Cook, 2017). The accessibility and prevalence of AI chatbots have raised significant ethical concerns regarding their involvement in the dissemination of disinformation. Numerous studies have highlighted the detrimental effects of disinformation on society. Research by Allcott and Gentzkow (2017) examined the impact of fake news during the 2016 U.S. presidential election, revealing that false stories circulated on social media platforms reached a substantial portion of the American population. This dissemination of misinformation not only influenced public opinions but also contributed to the polarization of political ideologies (Bakshy, Messing, & Adamic, 2015).

The use of AI chatbots as vehicles for disinformation dissemination intensifies these concerns. AI chatbots, powered by sophisticated algorithms, have the ability to interact with individuals and mimic human-like conversations. This capability allows them to engage in large-scale dissemination of false narratives, leading unsuspecting individuals to believe and share inaccurate information. Research conducted by Ferrara, Varol, Davis, Menczer, and Flammini (2016) examined the role of social bots in spreading misinformation on Twitter. Their findings revealed that a significant portion of political discussions on the platform were driven by automated accounts, which often disseminated false information. Similarly, Howard et al. (2018) explored the manipulation of political conversations on social media through the deployment of bots, highlighting the potential for these automated systems to amplify disinformation campaigns. The ethical implications of AI chatbots participating in disinformation campaigns are significant. The deliberate use of chatbots to spread falsehoods not only undermines the trustworthiness of information sources but also erodes public confidence in democratic processes and institutions (Guess, Nyhan, & Reifler, 2020). Moreover, the spread of disinformation through AI chatbots has the potential to exacerbate existing social divisions and fuel conflicts (Starbird, 2017).

To address this issue, researchers and policymakers have begun exploring solutions such as the development of AI-based tools capable of identifying and mitigating the influence of chatbot-driven disinformation campaigns (Ratkiewicz et al., 2011). Additionally, platforms and social media companies are implementing stricter policies and investing in AI-powered detection systems to combat the spread of disinformation (Ruchansky, Seo, & Liu, 2017).

Figure 2.1: Carnegie Mellon researchers used AI to transfer facial expressions from one video to another (cmu.edu/news "Deep Fakes and the Future of Video Content.")

2.2 Accountability and Responsibility:

AI systems like ChatGPT generate responses based on complex machine learning algorithms, AI technologies, such as OpenAI's ChatGPT, formulate responses utilizing intricate machine learning models (Radford et al., 2019), which complicate the process of attributing accountability when they spread disinformation. The determination of liability for harm induced by AI-aided disinformation represents a multifaceted issue that extends beyond mere technical concerns, enveloping legal and ethical dimensions (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). ChatGPT, like other AI models, operates on a Transformer-based language modeling technique, a form of deep learning (Vaswani et al., 2017). Despite its computational power, this approach lacks transparency in its response generation process, often labeled as the 'black box' problem (Castelvecchi, 2016). This opacity complicates tracing the decision-making pathways, rendering accountability arduous.

From a juridical standpoint, existing laws face challenges in addressing AI-induced disinformation due to the intangible nature of AI systems (Pagallo, 2013). Applying legal norms like defamation or incitement becomes nontrivial when an AI system disseminates disinformation (Schroeder, 2018). On the ethical plane, AI chatbots spreading disinformation unchecked can infringe principles such as truthfulness and fairness (Floridi & Cowls, 2019). Such uncontrolled dissemination can undermine trust in digital platforms, skew public opinion, and potentially lead to societal destabilization (Howard et al., 2018). Recognizing these challenges, there's an urgent call for regulatory frameworks and comprehensive ethical guidelines to mitigate potential harm and ensure beneficial AI development (Russell et al., 2015).

2.3 Privacy and Personal Data Misuse:

Artificial intelligence (AI) chatbots are capable of amassing an extensive array of data stemming from interactions with users. Nonetheless, this ability to gather vast amounts of data carries the risk of compromising user privacy if such data is not appropriately safeguarded or is exploited for unsanctioned purposes (Sebastian, G.,2023). For instance, a serious potential repercussion could be the utilization of this data for manipulative tactics or the spread of disinformation, echoing concerns raised by Zuboff (2019).

Additionally, the quality and integrity of the data used for training AI chatbots are of paramount importance. If the data set employed in the training process is subject to corruption or bias, there can be harmful consequences, such as the generation and dissemination of misleading or harmful information. Studies by Caliskan, Bryson, and Narayanan (2017) corroborate this notion, highlighting that AI algorithms can inadvertently learn and perpetuate systemic bias present in the training data.

2.4 Algorithmic Fairness:

The concept of algorithmic fairness and the necessity of vetting training data are particularly crucial in the context of Human Resource (HR) applications of AI chatbots like ChatGPT (Sebastian, G. ,2023). The notion of algorithmic fairness refers to the aim of ensuring that AI systems do not unduly favor or disadvantage any group based on characteristics such as gender, ethnicity, or age (Dwork et al., 2012). Given that HR decisions can have significant impacts on individuals' careers and livelihoods, it's vital that these decisions are made fairly and without bias, as these often involve making decisions that can impact individuals' professional lives, such as hiring, promotion, and compensation decisions.

However, as researchers have highlighted, AI systems can inadvertently learn to replicate or even amplify the biases present in their training data (Barocas & Selbst, 2016) 2. For eg, if an AI chatbot like ChatGPT is trained on data where certain professions are predominantly associated with a particular gender, it might propagate this bias in its interactions with users. Several recent studies have demonstrated that LLMs, such as GPT-3, have a persistent bias against gender (L. Lucy and D. Bamman, 2021) and religions (A. Abid, M. Farooqi, and J. Zou, 2021). There could also be monolingual bias in multilingualism that can occur in language models (Z. Talat, A. et al., 2022) To overcome this, it is crucial to ensure that the training data contains a substantial proportion of diverse, high-quality corpora from various languages and cultures.

2.5 The prompt injection:

This includes input as data that is deliberately introduced into the model’s input with the intention of causing it to malfunction. To address this vulnerability, it is crucial to conduct exhaustive testing on a wide variety of

inputs and ensure that the model can accurately recognize and reject inputs that are different from the semantic and syntactic patterns of the input it was trained on. Additionally, it is essential to establish robust monitoring methods to detect any malicious use of the model and implement necessary security measures to prevent such malicious intent (Zhuo, T. Y. et al, 2023).

3.0 Survey Results:

The survey to understand the ethical implications of ChatGPT and other AI Chatbots and proposed steps for disinformation regulation was conducted on Amazon's Mechanical Turk (MTurk) platform from May 7 to May 21, received 201 responses. The significance of this survey data can be interpreted in terms of statistical significance. The number of responses (201) exceeds the minimum sample size required for a survey to attain statistical significance, typically set at around 30 responses, according to the Central Limit Theorem (Lumley, Diehr, Emerson, & Chen, 2002). See Table 3.1 below and Figures 3.4 for further details.

Summary of survey results	**Responses**
1. How aware are you about using AI-based Chatbots like ChatGPT (on a scale of 1-5)	
- 3	18 (9%)
- 4	120 (59.7%)
- 5	62 (30.8%)
2. Geography of the survey respondents	
- North America	93 (46.7%)
- South America	77 (38.7%)
- Asia Pacific	15 (7.5%)
- Africa	06 (3%)
- Europe	08 (4%)
3. Please select your gender and age group	
- Male 28 – 45 years	119 (59.8%)
- Male 46+ years	23 (11.6%)
- Female 28 – 45 years	52 (26.1%)
- Female 46+ years	05 (2.5%)
4. To what extent do you believe that AI chatbots, such as ChatGPT, have the potential to spread disinformation or misinformation?	
- 1	02 (1%)
- 2	08 (4%)
- 3	29 (14.4%)
- 4	129 (64.2%)
- 5	33 (16.4%)
5. **How important do you think it is for AI developers and companies to prioritize addressing ethical implications and minimizing the spread of disinformation in AI chatbots?**	
- 1	01 (0.5%)
- 2	07 (3.5%)
- 3	28 (14.1%)
- 4	106 (53.5%)
- 5	56 (28.3%)
6. Do you think that AI chatbots should be regulated by an external organization to prevent disinformation propagation?	
- Yes	173 (86.1%)
- No	14 (7%)
- Not Sure	14 (7%)
7. What level of responsibility should AI developers have in mitigating the spread of disinformation through their chatbot systems?	
- 1	1 (0.5%)
- 2	4 (02%)
- 3	29 (14.5%)
- 4	128 (64%)
- 5	38 (19%)

8. How much transparency should be expected from AI developers regarding their efforts to minimize disinformation in their chatbots?	
- 1	1 (0.5%)
- 2	4 (2%)
- 3	33 (16.8%)
- 4	98 (49.7%)
- 5	61 (31%)
9. In your opinion, which of the following stakeholders should be primarily responsible for preventing AI chatbots from spreading disinformation?	
- AI Developers	139 (69.8%)
- Government Agencies	32 (16.1%)
- Independent Organizations	26 (13.1%)
- End Users	02 (01%)
10. Are you aware of any existing ethical guidelines or regulations for AI chatbot development? If so, do you think they are sufficient to address the issue of disinformation?	
- Yes	163 (83.2%)
- No	20 (10.2%)
- Not Sure	13 (6.6%)
11. Would you support the implementation of a mandatory rating or certification system for AI chatbots based on their potential to spread disinformation?	
- Yes	167 (84.3%)
- No	24 (12.1%)
- Not Sure	07 (3.5%)
12. In your opinion, which of the following stakeholders should be primarily responsible for preventing AI chatbots from spreading disinformation?	
- Increase awareness among everyone in the organization	11 (5.56%)
- Develop and implement standards for AI systems to ensure accuracy and fairness.	14 (07%)
- Incorporate a feedback loop into AI systems so that users can comment on information accuracy and fairness.	25 (12.6%)
- Use automated tools to detect and flag false information	36 (18.18%)
- Use data and algorithms that are transparent and explainable.	41 (20.71%)
- Establish independent oversight boards to ensure compliance with ethical guidelines.	30 (15.15%)
- Ensure data for AI model creation is accurate and up-to-date	41 (20.71%)

Table 3.1: Survey results on ethical issues of AI-based Chatbots like ChatGPT

3.1 Regulation of Disinformation Propagation - Transparency and Explainability

Regulation to ensure transparency and explainability of AI chatbots is vital in mitigating the propagation of disinformation. This is particularly significant given that AI chatbots, like ChatGPT, are becoming increasingly utilized across various sectors, including news and social media platforms, where the spread of disinformation could have far-reaching societal consequences.

3.1.1 Transparency: Transparency in AI systems refers to the ability to see clearly into how the system operates. This can help users, regulators, and the public understand the methods and principles guiding the AI's behavior. For chatbots, transparency could involve clarity about the sources of training data, the nature of the algorithm used, and any biases inherent in the system (Holstein, Wortman Vaughan, Daumé III, Dudik, & Wallach, 2019).

3.1.2 Vetting Training Data: To mitigate these risks, it's crucial to vet the training data used for AI chatbots. This involves carefully examining the data to ensure it's as representative, unbiased, and fair as possible. Firstly, potential sources of bias in the data must be identified. This could involve pinpointing underrepresented groups or recognizing societal biases that could be present in the data (Gebru et al., 2018). Secondly, strategies should be applied to mitigate these biases. This could range from oversampling underrepresented groups to applying bias-correction algorithms (Bolukbasi et al., 2016). Finally, the modified dataset must be tested to ensure that it doesn't lead to biased outcomes when used for training an AI model. This involves quantitative evaluations,

like measuring disparity in outcomes for different groups, and qualitative evaluations, like having human reviewers examine the AI's decisions. Further it also needs to be made sure that the data is not outdated.

3.1.3 ***Explainability*** refers to the ability to provide understandable explanations for a chatbot's decisions or recommendations. It can help users make informed assessments about the reliability of the information provided by the chatbot (Ribeiro, Singh, & Guestrin, 2016). For eg, a language technology that analyzes curricula vitae for recruitment or career guidance based on historical data may be less likely recommend historically discriminated groups to recruiters or more likely to offer lower-paying occupations to marginalized groups. To prevent this, it is essential to ensure that the training data is diverse and representative of the population for which it will be used, and to actively discover and eradicate any potential biases in the data (Zhuo, T. Y., et. al, 2023).

3.1.4 ***Periodic Audits:*** external audits and third-party evaluations of AI chatbot systems could be mandated by regulatory authorities to ensure adherence to transparency and explainability principles, thus further limiting disinformation propagation (Bishop, 2021). Moreover, developers could be obligated to implement design features that discourage the spread of false information. For instance, an AI system could be required to flag when it's generating content in areas where it has limited or potentially biased training data, which could reduce the risk of disinformation being generated and spread.

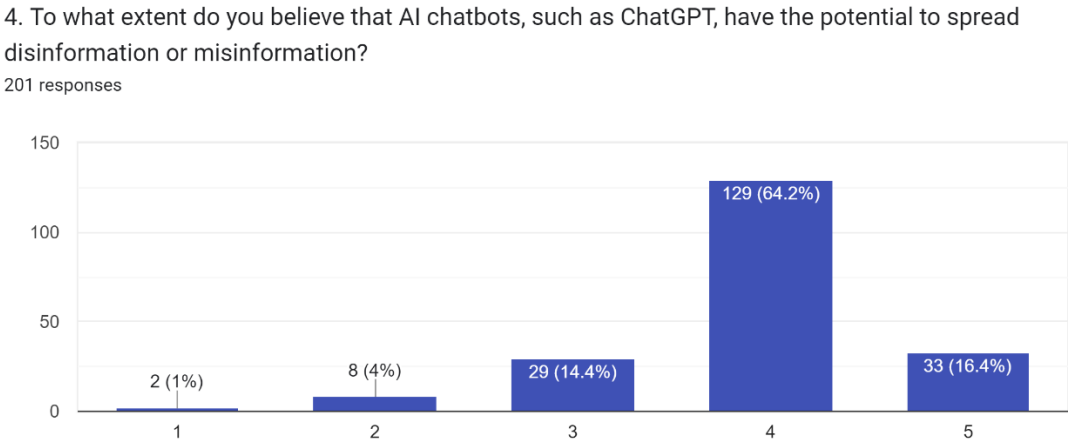

Figure 3.1: Survey results on what extent do AI chatbots have potential to spread disinformation

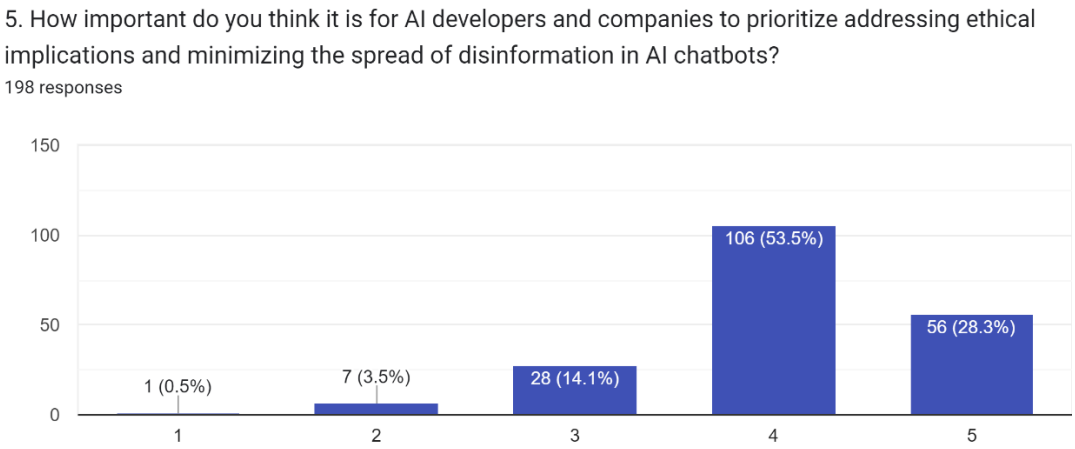

Figure 3.2: Survey results on the significance of AI developers and companies to prioritize ethical implications of AI chatbots

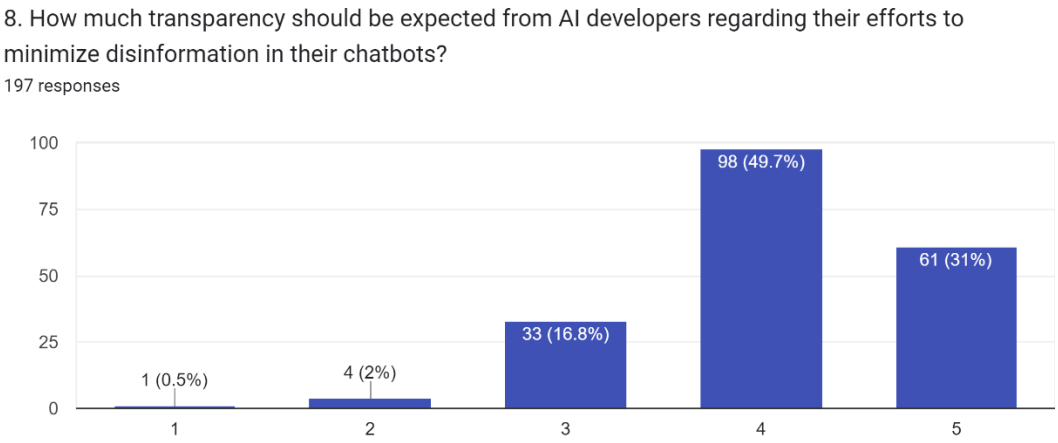

Figure 3.3: Survey results on the transparency from AI developers regarding their efforts to minimize disinformation in chatbots

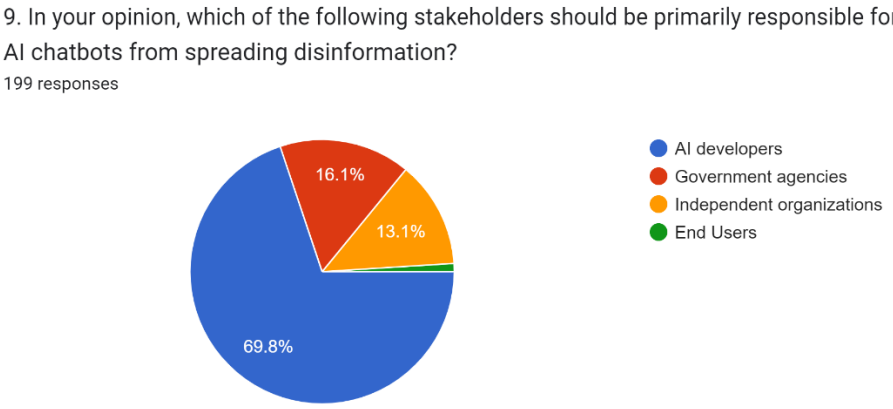

Figure 3.4: Survey results on which stakeholder should be primarily responsible for preventing AI chatbots from spreading disinformation

3.2 Accountability Measures

The propagation of disinformation by AI chatbots poses significant challenges. To address these, it is crucial that regulatory frameworks articulate clear lines of accountability for AI-generated disinformation. This is an emerging area in AI ethics and law, which can be supported by scholarly research in the field.

3.2.1 ***Accountability*** refers to the idea that entities (individuals, organizations, or systems) should be held responsible for their actions, particularly when those actions have significant consequences for others. In the context of AI, accountability can be complex due to the multi-layered nature of AI development and deployment (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). Regulations can play a pivotal role in establishing accountability for AI-generated disinformation. These could take the form of laws and policies that hold different stakeholders, such as developers, deployers, or users of AI chatbots, accountable depending on the specific circumstances (Schiff, Biddle, Borenstein, & Laas, 2020).

3.2.2 ***Regulatory frameworks:*** Regulation can play a crucial role in promoting explainability and transparency. For instance, guidelines could require AI developers to provide explanations of their systems' decision-making processes, either in a technical form for expert scrutiny or in a more user-friendly form for laypeople (Guidotti et al., 2018). Other measures include requiring developers to maintain thorough documentation of their design and training processes. This could enable better scrutiny of potential sources of bias or vulnerability to misuse in AI chatbots (Gebru et al., 2018). Deployers, who are typically companies or

organizations using the AI chatbot, could be held accountable for regularly auditing the chatbot's performance to ensure that it doesn't propagate disinformation, and for swiftly addressing any issues that arise. Users who deliberately manipulate AI chatbots to spread disinformation could also be held accountable under such regulations. This could involve the application of existing laws relating to disinformation and manipulation of public discourse, or the development of new laws specifically tailored to the unique challenges posed by AI technologies (Chesney & Citron, 2018). Such a multifaceted approach to accountability can help to deter misuse of AI chatbots for disinformation purposes, and ensure that those who do misuse these technologies are held responsible.

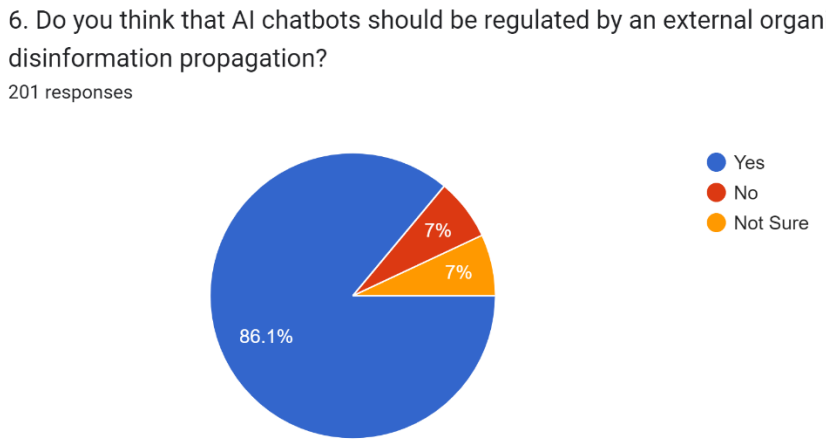

Figure 3.5: Survey results on if AI chatbots should be regulated by an external organization

3.3 Data Privacy Regulations

AI chatbots like ChatGPT process a substantial amount of user data, raising significant data privacy concerns. Stricter data privacy regulations can potentially reduce the misuse of personal data by these chatbots, playing a vital role in ensuring that user data is not used to generate or propagate disinformation. Data privacy concerns arise in AI chatbots due to the significant amount of data that is processed to make these systems work efficiently. In certain cases, sensitive personal data may be unintentionally included in conversations with the chatbot, increasing the risks associated with potential data misuse (Custers & Ursic, 2020). Furthermore, AI chatbots can be used for nefarious purposes, such as spreading disinformation tailored to individual users based on their personal data. This represents a significant threat to user privacy and information integrity.

3.3.1 GDPR and other European Data directives: Stricter data privacy regulations can mitigate these risks by stipulating how user data should be collected, processed, stored, and used. Regulations such as the General Data Protection Regulation (GDPR) in the European Union already lay out strict rules for data privacy, which can serve as a starting point for developing regulations targeted at AI chatbots (Voigt & Von dem Bussche, 2017). Such regulations could potentially mandate that chatbot developers and deployers obtain explicit consent from users before processing their data. Additionally, they could require that developers implement robust data anonymization and encryption techniques to protect user data (Danezis & Gürses, 2010).

3.3.2 Data privacy regulations can also play a role in mitigating disinformation. By limiting how personal data can be used, these regulations can make it more difficult for AI chatbots to generate disinformation tailored to individual users. Furthermore, if users have more control over their data, they may be more able to prevent their data from being used to spread disinformation.

4.0 Monitoring and Reporting Mechanisms

Regulation The adoption of robust monitoring and reporting mechanisms is crucial in regulating AI chatbots such as ChatGPT. These measures play a significant role in identifying and mitigating the spread of disinformation, helping to ensure the reliability and integrity of information generated by these systems.

4.1. Third-Party Audits: Third-party audits of AI systems can be an effective mechanism to monitor the dissemination of disinformation. Independent audits can provide an unbiased evaluation of a chatbot's performance, highlighting potential vulnerabilities to disinformation propagation (Knight, 2020) 1. These audits could assess factors such as the transparency of the AI's decision-making process, its susceptibility to manipulation, and its ability to detect and reject false information. Regular audits can also ensure continued adherence to transparency and explainability guidelines, which are vital for trustworthiness in AI (European Commission's High-Level Expert Group on AI, 2019).

4.2. Public Reporting of Disinformation Incidents: Public reporting of disinformation incidents is another effective monitoring mechanism. A reporting system that requires developers or deployers of AI chatbots to publicly disclose any incidents of disinformation can promote transparency and accountability (Tschantz, Datta, & Wing, 2012) 3. Such transparency can enable users, regulators, and the public to better understand the potential risks associated with using these systems, and inform decision-making about their use.

4.3. User-Driven Reporting Systems: Lastly, implementing systems for users to report suspected disinformation is also crucial. Given the vast amount of content generated by AI chatbots, user involvement can be an invaluable resource in detecting and flagging disinformation (Cummings, 2020) 4. These reports can then be used to refine the AI system and improve its ability to detect and prevent disinformation in the future.

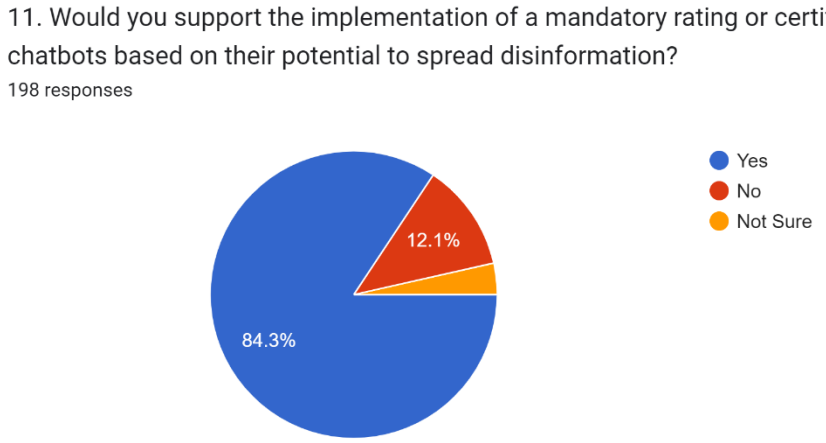

Figure 4.1: Survey results on the need for implementation of a mandatory rating or certification system

5.0 Conclusion

The rapid growth and increasing influence of AI chatbots, exemplified by the emergence of advanced models like ChatGPT, have undeniably provided numerous benefits in various domains. However, alongside these advantages, the widespread adoption and utilization of AI chatbots have also given rise to a host of ethical challenges, particularly concerning the propagation of disinformation. The ability of AI chatbots to generate human-like responses and engage in real conversations raises concerns about the potential spread of false information, manipulation of public opinion, and the erosion of trust in digital spaces. As AI technology continues to advance at a rapid pace, it is crucial to recognize that the ethical implications and regulatory requirements surrounding AI chatbots and their impact on disinformation will also evolve. Efforts to address these challenges must be ongoing and must involve multidisciplinary collaboration between researchers, policymakers, technologists, ethicists, and other stakeholders. Such collaborative engagement will help ensure that the development, deployment, and use of AI chatbots align with societal values and serve as tools for the betterment of society, rather than causing harm.

To effectively navigate the complex landscape of AI chatbots and disinformation, regulatory frameworks need to be established or adapted to address the unique challenges posed by these technologies. These frameworks should encompass considerations such as transparency, accountability, data privacy, algorithmic bias, and the detection and mitigation of disinformation. Striking the right balance between fostering innovation and safeguarding against potential harms is paramount. Moreover, as the field of AI continues to progress, ongoing

research and technological advancements must be accompanied by continuous evaluation of the ethical implications of AI chatbots. This evaluation should involve rigorous assessment of their impact on society, including the potential risks of disinformation propagation, the psychological effects on users, and the broader implications for democratic processes and public discourse. Regular monitoring and assessment will enable prompt identification of emerging issues and the implementation of necessary safeguards and regulations.

In conclusion, the advent of AI chatbots, exemplified by the remarkable capabilities of ChatGPT, has ushered in transformative possibilities. However, it is vital to recognize and address the ethical challenges associated with these technologies, particularly their potential role in disinformation propagation. Ongoing multidisciplinary engagement, comprehensive regulatory frameworks, and continuous evaluation are essential components in ensuring that AI chatbots are wielded responsibly and ethically and that they ultimately contribute to the betterment of society."

5.1 Future study scope:

Scope for the future study includes studying the scope of embedding risk and controls by default within the large language models Glorin, S. (2020). Another proposed study includes extending the ethical implications of large language models to other technology components such as Metaverse, Glorin, S. (2023), and mature business intelligence and analytics in organizations (George, A. et. al 2018, George, A. et. al 2020).

5.2 Statements and Declarations

Competing Interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data Availability: The generated during and/or analyzed during the current study are available in the Glorin repository, 10.6084/m9.figshare.23258006

References

[1]. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236.

[2]. Abid, M. Farooqi, and J. Zou, “Persistent antimuslim bias in large language models,” in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 298–306.

[3]. Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130-1132.

[4]. Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104(3), 671-732.

[5]. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

[6]. Caliskan, A., Bryson, J.J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186. DOI: 10.1126/science.aal4230

[7]. Castelvecchi, D. (2016). Can we open the black box of AI? Nature News, 538(7623), 20-23.

[8]. Chesney, R., & Citron, D. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.

[9]. Cummings, C. (2020). How Users Can Report Disinformation and Why They Don’t: A Study on Reporting Tools. The Partnership on AI. https://www.partnershiponai.org/reporting-disinformation/

[10]. Custers, B., & Ursic, H. (2020). Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. International Data Privacy Law, 10(1), 32-46. DOI: 10.1093/idpl/ipz020

[11]. Danezis, G., & Gürses, S. (2010). A critical review of 10 years of Privacy Technology. Proceedings of Surveillance Cultures: A Global Surveillance Society?.

[12]. Derner, E., & Batistič, K. (2023). Beyond the Safeguards: Exploring the Security Risks of ChatGPT. arXiv preprint arXiv:2305.08005.

[13]. "Deep Fakes and the Future of Video Content." Carnegie Mellon University News, 21 September 2018, https://www.cmu.edu/news/stories/archives/2018/september/deep-fakes-video-content.html.

[14]. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12. doi: 10.1145/2090236.2090255

[15]. European Commission's High-Level Expert Group on AI. (2019). Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[16]. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96-104.

[17]. Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review.

[18]. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for Datasets. ArXiv, abs/1803.09010.

[19]. George, Amrita, Kurt Schmitz, and Veda C. Storey. "A framework for building mature business intelligence and analytics in organizations." Journal of Database Management (JDM) 31.3 (2020): 14-39.

[20]. George, Amrita, Kurt Schmitz, and Veda Storey. "The BI&A system: building matured business intelligence in organizations." Academy of Management Global Proceedings 2018 (2018): 144.

[21]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey Of Methods For Explaining Black Box.

[22]. Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. Nature Human Behaviour, 4(5), 472-480.

[23]. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper No.: 600. DOI: 10.1145/3290605.3300830

[24]. Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). The IRA, Social Media and Political Polarization in the United States, 2012-2018. Oxford University.

[25]. Howard, P. N., Kollanyi, B., & Bradshaw, S. (2018). Bots, #Strongerin, and #Brexit: Computational propaganda during the UK-EU referendum. SSRN Electronic Journal.

[26]. Knight, W. (2020). AI is wrestling with a replication crisis. WIRED. https://www.wired.com/story/ai-wrestling-with-replication-crisis/

[27]. L. Lucy and D. Bamman, "Gender and representation bias in gpt-3 generated stories," in Proceedings of the Third Workshop on Narrative Understanding, 2021, pp. 48–55.

[28]. L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh et al., "Ethical and social risks of harm from language models," arXiv preprint arXiv:2112.04359, 2021.

[29]. Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. Journal of Applied Research in Memory and Cognition, 6(4), 353-369.

[30]. Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The Importance of the Normality Assumption in Large Public Health Data Sets. Annual review of public health, 23, 151-169. DOI: 10.1146/annurev.publhealth.23.100901.140546

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

[32]. Pagallo, U. (2013). The laws of robots: crimes, contracts, and torts. Springer Science & Business Media.

[33]. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.

[34]. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In Proceedings of the 20th international conference companion on World wide web (pp. 249-252). ACM.

[35]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. DOI: 10.1145/2939672.2939778

[36]. Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. arXiv preprint arXiv:1703.09931.

[37]. Russell, S., Dewey, D., & Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. Ai Magazine, 36(4), 105-114.

[38]. Schroeder, T. (2018). Speech, lies, and problematic causation. Ethics and Information Technology, 20(3), 195-203.

[39]. Sebastian, Glorin. "Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information."

[40]. Sebastian, Glorin. "Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk?: An Exploratory Study." International Journal of Security and Privacy in Pervasive Computing (IJSPPC) 15.1 (2023): 1-11.

[41]. Sebastian, Glorin. "A Descriptive Study on Metaverse: Cybersecurity Risks, Controls, and Regulatory Framework." International Journal of Security and Privacy in Pervasive Computing (IJSPPC) 15.1 (2023): 1-14.

[42]. Sebastian, Glorin. "Evolution of the role of risk and controls team in an ERP Implementation." IJMPERD, ISSN (P) (2020): 2249-6890.

[43]. Sebastian, G. (2023). Hello! This is your new HR Assistant, ChatGPT! Impact of AI Chatbots on Human Resources: A Transformative Analysis. DOI: 10.13140/RG.2.2.21668.86405.

[44]. Sebastian, Shaliet Rose, and Bichu P. Babu. "Impact of metaverse in health care: a study from the care giver's perspective." International Journal of Community Medicine and Public Health 9.12 (2022): 4613.

[45]. Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1-23.

[46]. Tschantz, M. C., Datta, A., & Wing, J. M. (2012). Formalizing and Enforcing Purpose Restrictions in Privacy Policies. 2012 IEEE Symposium on Security and Privacy, 176-190. DOI: 10.1109/SP.2012.18

[47]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 5998-6008.

[48]. Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). Springer Publishing.

[49]. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books.

[50]. Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring ai ethics of chatgpt: A diagnostic analysis. arXiv preprint arXiv:2301.12867.

[51]. Z. Talat, A. N´ev´eol, S. Biderman, M. Clinciu, M. Dey, S. Longpre, S. Luccioni, M. Masoud, M. Mitchell, D. Radev et al., "You reap what you sow: On the challenges of bias evaluation under multilingual settings," in Proceedings of BigScience Episode# 5– Workshop on Challenges & Perspectives in Creating Large Language Models, 2022, pp. 26–41.