On the Ethical Considerations of Generative Agents

N'yoma Diamond
Department of Computer Science and Technology
University of Cambridge
Cambridge, England
bad35@cantab.ac.uk

Soumya Banerjee
Department of Computer Science and Technology
University of Cambridge
Cambridge, England
sb2333@cam.ac.uk

Abstract

The Generative Agents framework recently developed by Park et al. has enabled numerous new technical solutions and problem-solving approaches. Academic and industrial interest in generative agents has been explosive as a result of the effectiveness of generative agents toward emulating human behaviour. However, it is necessary to consider the ethical challenges and concerns posed by this technique and its usage. In this position paper, we discuss the extant literature that evaluate the ethical considerations regarding generative agents and similar generative tools, and identify additional concerns of significant importance. We also suggest guidelines and necessary future research on how to mitigate some of the ethical issues and systemic risks associated with generative agents.

1 Introduction

The seminal work by Park et al. [1] introduced the Generative Agents framework for simulating human behaviour using generative language models. Generative agents have the ability to operate independently and creatively, making decisions to reach a goal with minimal user input. To do this, each agent maintains a distinct and dedicated set of memories and behavioural parameters that are provided to a generative language model as needed. By recording memories, prompting reflection on them, planning future actions, and reacting to short-term observations, generative agents can produce realistic emulations of human behaviour and decision-making [1]. As a result, generative agents have massive potential for widespread adoption and may revolutionise many modern problems. However, this introduces a number of critical ethical challenges regarding their research, development, and application. In this position paper, we examine the existing literature on the ethical aspects of generative agents, highlight a handful of critical ethical concerns, and discuss recommendations for their mitigation. We note that the concerns posed in this work are not exhaustive or entirely exclusive to generative agents, but are believed to be especially important to discuss in this specific context.

2 Existing Literature

Bail [2] provides a detailed qualitative analysis and discussion of the potential benefits of generative artificial intelligence (GAI) and generative agents towards social science. With respect to ethical considerations, Bail [2] primarily focuses on GAI over generative agents specifically. Bail [2]

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

identifies that GAI tools are highly limited by their reliance on human-generated knowledge and are thereby vulnerable to exhibit human biases. They also question whether GAI can produce high-quality realistic and accurate research which is truly actionable, particularly informed by many existing GAI technologies' propensity to hallucinate and produce misinformation. Finally, Bail [2] raises concerns over the reproducibility of research using GAI due to its inherent stochasticity and the massive quantity of non-peer-reviewed GAI-based literature being produced, potentially harmfully contributing to ongoing replication crises [3–6]. However, Bail [2] also identifies the critical opportunities for open social science research posed by GAI, such as through the development and improvement of open-source GAI tools and infrastructure, citing precedent from the (former) open sharing of research data generated by many social media companies.

Lazar [7] considers a narrative rhetoric approach towards analysing the potential societal impacts of generative agents. In their work, Lazar [7] identifies prior challenges produced by GAI, their relevance towards recent developments in GAI, and existing approaches' failure to overcome some of these hurdles. They also discuss specific applications of generative agents and the ethical questions and concerns they face from the perspective of societal impact: Generative agent-based companions can provide social engagement, assist in day-to-day tasks, improve recommender and information retrieval systems, and serve as "universal intermediaries" by consolidating and automating useful tasks and jobs. However, this may cause users to develop unhealthy parasocial relationships and dependencies, displace human labour, commodify human attention, and worsen user privacy and security [7]. Notably, while Lazar [7] adequately identifies these numerous challenges, they do some from a primarily rhetorical standpoint. As a result, they neglect to suggest means to address many of the stated concerns.

Chan et al. [8] provide an overview and analysis of future harms of agentic algorithmic systems more broadly. By focusing on agentic algorithmic systems as a broad category (including generative agents, among other technologies), Chan et al. [8] justify the need to anticipate future harms of such systems, identify specific examples, and propose approaches and areas of future work to explore and implement to prevent the identified harms before they occur. Notably, while they do make note of GAI-based tools for specific examples, they are not a specific focus of the work. As a result, many risks and challenges specific to agentic GAI-based tools, like generative agents, are ignored.

Anwar et al. [9] and Gabriel et al. [10] provide comprehensive and in-depth discussion of the technical challenges and ethical risks of GAI-based systems, respectively. Both works specifically consider generative agents or similar systems, making them the most directly related extant literature to this work. Anwar et al. [9] primarily focus on the technical challenges of GAI-based systems and how flaws in their design, training, tuning, and implementation may lead to critical problems. Specifically, in their discussion about agentic LLMs and multi-agent systems, Anwar et al. [9] identify risks resulting from the combination of generative natural language models and agentic behaviour, such as difficulty ensuring desired behaviour, providing robust oversight, and ensuring safe behaviour when afforded access to external abilities and services. The authors also identify risks associated with the emergent behaviours of multi-agent systems and the lack of extant research into understanding them directly. Based on these and other identified challenges, Anwar et al. [9] discuss a series of sociotechnical challenges, many of which have ethical implications in line with those discussed in this work. Gabriel et al. [10] specifically focus on what they refer to as "AI assistants", which are GAI-based agents capable of independent and autonomous planning and action on behalf of a user via text-based instruction. This case is a specific example within the broader category of generative agents, however it serves as an effective case-study towards identifying the risks and ethical concerns of generative agents. Using this lens, Gabriel et al. [10] identify a series of significant personal, socioeconomic, and environmental concerns posed by AI assistants, many of which we discuss in this work.

3 Ethical Concerns of Generative Agents

Anthropomorphisation and Misunderstanding of Experimental Results

Anthropomorphisation can serve as an effective means to discuss and rationalize the behaviour of AI tools, such as generative agents. However, generative agents and other GAI tools as they exist today do not actually possess any level of consciousness and are incapable of producing true emotion or intention. Thus, a problematic disconnect can develop between how AI tools are understood,

discussed, and utilised as compared to the reality of their capabilities. This poses a significant ethical concern for the usage and understanding of generative agents with respect to human behaviour. Present literature provides substantial discussion of anthropomorphisation of generative agents and similar tools towards users in social applications (such as chatbots) [7, 10, 11]. However, no extant research has been identified directly analysing the impacts and risks associated with how results generated using generative agents may be erroneously interpreted due to excessive anthropomorphisation. That is, prescribing undue anthropomorphic characteristics to generative agents risks understating its purpose as an emulator of human behaviour, and not a direct model of how humans actually behave.

A critical benefit of generative agents is the potential to serve as an effective tool to perform human-like actions where humans cannot be used, such as to emulate and understand human behaviour or complex human-interaction systems. However, excessive anthropomorphisation risks critically misunderstanding the results of such experiments and creating misinformation. That is, any result produced using generative agents is only descriptive of the behaviour of the implemented framework and its underlying generative model. The behaviours of generative agents and language models are not causally linked to human behaviour patterns. To ensure the quality and accuracy of research conducted using generative agents, conscious effort must be made to avoid misattributing the characteristics of generative agents towards the behaviour of real humans. In addition, more research must be conducted to understand the risks associated with such distorted interpretations.

Creation of Parasocial Relationships

Anthropomorphisation also risks the creation of parasocial relationships between AI and its users [7, 10]. This is particularly true for generative agents, which may be individualised and/or physically embodied in future applications, such as explored by Gabriel et al. [10]. The structure of generative agents lend themselves towards the creation of individualised agents which non-technical users may develop relationships. Individualised or embodied generative agents may be an effective tool toward solving day-to-day problems—much like existing virtual assistants such as Siri and Google Assistant. However, the persona, memory, and reflection characteristics of generative agents allow them to be perceived as more similar to humans than existing virtual assistants despite not being fundamentally any more "intelligent". As a result, the users of such agents risk harmfully anthropomorphising GAI and developing harmful relationships or attachments to these tools.

While Gabriel et al. [10] discuss ways to mitigate risks induced by parasocial relationships, they mainly focus on post hoc utilitarian and material approaches to optimise reliable agent performance in cases of human-AI relationships, rather than avoiding them altogether. Park et al. [1] and Abercrombie et al. [11] suggest that generative agents and other GAI tools should be designed to explicitly state their nature as generative models and directly avoid anthropomorphic language and behaviour. However, little consideration has yet to be given to how this may be reliably implemented and the effectiveness of such approaches. In particular, many of these methods burden the user with identifying and mitigating parasocial relationships. Such approaches may prove unhelpful or even counterproductive if users ignore, misinterpret, or intentionally circumvent the provided warnings or guardrails [9]. Further, anthropomorphic characteristics are not objectively identifiable and the boundaries distinguishing anthropomorphic behaviour are becoming exceedingly unclear. As a result, approaches to reliably avoid anthropomorphic behaviours will be extremely difficult to develop.

As a simple example, the implementation provided by Park et al. [1] for their generative agent simulation framework defers to the underlying generative model to identify unsafe behaviour. Upon query, the generative model is prompted to score how much a user's prompt anthropomorphizes the agent. If the score is greater than a critical threshold, then the user is told that attributing human agency to generative agents is inappropriate and their query is rejected. This aims to ensure that the framework is being used safely. However, this method is very simplistic and relies on the untested assumption that the underlying model is capable of accurately identifying harmful anthropomorphisation. In reality, the ability to identify such characteristics will vary greatly between models and is not provably accurate. In addition, this system provides no guardrails to prevent the user from ignoring or intentionally circumventing these checks.

Excessive Trust and Insufficient Scepticism

Similar to concerns regarding the misattribution of human characteristics, over-reliance and overconfidence in generative agents may result in the unintentional spread of misinformation. Specifically, users may prescribe undue trustworthiness to generative agents [9, 10, 12, 13]. Intuitively, generative agents should be more capable than existing technologies (such as AI chatbots or virtual assistants) due to their memory and reflection modules, as these modules allow generative agents to better recall information and develop critical insights [1]. However, generative agents are still just an extension of existing language models and thus are prone to the same mistakes and errors. Such errors include hallucinations, poisoning, failure to recall available information, or recapitulating and reinforcing endemic biases. Further, the stochasticity of generative language models can induce unreliable and/or inaccurate behaviour from generative agents [2, 14].

A lack of critical analysis and scepticism of responses produced by generative agents runs the risk of unintentionally spreading misinformation and reinforcing harmful biases [9, 10, 12, 13]. Gabriel et al. [10] specifically discuss the angle of (misplaced) trust and misinformation associated with AI assistants. In particular, they identify varying types of trust that a user may have in GAI-based systems like generative agents, such as overestimating the competence of an agent or the quality of its alignment. They also point out that undue trust makes people highly vulnerable to misinformation, manipulation, and ideological entrenchment—dynamics highly similar to those observed in humans, and potentially even more concerning if generative agents face widespread adoption. To address these risks, Gabriel et al. [10] collate a handful of technical design and policy-based proposals, such as identifying and indicating uncertainty, minimising (un)necessary complexity, and improving technical transparency and public understanding of GAI systems. However, many of these approaches require severely limiting desirable agent functionality or leveraging external response analysis methods which are vulnerable to circumvention. Thus, they may only serve as stop-gap solutions. As a result, there is still a significant need for further research and development of techniques that can inherently mitigate these concerns.

Ideally, internal technical mechanisms should be developed to provably ensure that provided information is true and impartial, or otherwise marked as uncertain or potentially biased. To this end, some researchers have suggested that retrieval-augmented generation (RAG) techniques may provide a technical solution to mitigate generative agents spreading or unintentionally creating misinformation by leveraging access to external knowledge resources during inference [15–17]. However, we were unable to identify any literature applying these techniques to memory-/reflection-enabled generative agents as proposed by Park et al. [1]. Further, RAG simply provides more reliable access to information and does not inherently prevent them from producing misinformation. It is possible for RAG-based agents to produce erroneous inferences that are only partially informed by retrieved information. That is, RAG-enabled generative agents may draw incorrect conclusions when provided with incomplete or highly complex knowledge. In addition, information retrieval can itself unintentionally provide the model with misinformation under certain circumstances, such as via design error or scope misalignment. This can cause even greater harm due to the elevated credibility that would likely be attributed to RAG-enabled generative agents [15, 17].

Usage by Malicious Actors

The prevalence of automated bots has become a key driver in the spread of misinformation online, significantly harming access to trustworthy and accurate information [9, 10, 18–22]. Simultaneously, GAI tools such as ChatGPT are beginning to be applied toward the creation of automated scams and phishing attacks [9, 10, 23, 24]. As automated means for conducting malicious acts become more prevalent, generative agents may be particularly susceptible to misuse. Malicious actors can leverage generative agents as automated tools to spread disinformation, execute scams, or conduct cyberattacks. The distinct memory and reflection characteristics of generative agents make them capable of performing malicious actions (such as spreading misinformation or conducting scams) with greater realism and effectiveness than existing technologies. Thus, malicious generative agents may be better at deceiving humans and avoiding automated detection than existing approaches. Gabriel et al. [10] briefly discuss the types of security vulnerabilities GAI systems introduce to users, and how malefactors' abilities may be enhanced by GAI systems leveraging memory, planning, and reflections capabilities like those proposed by Park et al. [1]. However, they do not consider specific concerns posed by automated generative agents.

Further, malicious generative agents may perform social engineering attacks, impersonate friends, relatives, or officials, and conduct other cyberattacks that were previously only possible by humans [25, 26]. To this end, future work should consider developing techniques to detect the usage of automated generative agents [10]. However, such work is easier said than done, as automatically and accurately distinguishing and mitigating malicious behaviour is exceptionally difficult. In practice, it may not be effective to specifically target GAI in lieu of broader misinformation, scam, or attack detection. Thus, additional practical solutions also need to come from legislating acceptable development and usage of generative agents to suppress misuse [9, 10].

Vulnerability to Hijacking

In addition to direct usage by malicious actors, developers of generative agents must be wary of their vulnerability to hijacking or jailbreaking. GAI tools are prime targets for attacks that aim to derail system behaviour. This is due to their usage in a wide variety of applications, the transferability of attacks between implementations, and the difficulty of developing effective behavioural guardrails [27, 28]. As with previously discussed concerns, generative agents' memory and reflection capabilities make them highly desirable and effective for an increasing range of tasks, raising direct concerns about hijacking as they see greater adoption. Hijacked generative agents may be difficult to recover, have access to more sensitive information and actions than other GAI tools, and potentially automatically hijack other agents or systems like a worm virus. This is especially concerning as substantial research has already been conducted towards developing techniques to compromise generative language models [27–31]. Notably, the desire to hijack GAI systems is not exclusive to malicious actors, as normal users may also directly benefit from hijacking (or "jailbreaking") automated AI-based tools [32–34]. Thus, hijacking and jailbreaking serve both as a threat vector for external actors to harm users of generative agents and for users to circumvent their safety features [9, 10].

In response to these concerns, many authors suggest a need to improve approaches to detect, understand, and mitigate agent hijacking [9, 10]. However, we believe a simpler approach should be considered with greater interest; critical assessment of the usage of generative agents altogether. Given the difficulty of detecting and preventing model hijacking and our as-yet lacking understanding of these problems [9], developers must be wary of where, when, how, and if they should implement generative agents at all. In particular, generative agents should not be utilised in contexts where their hijacking may enable significant threat vectors, such as those suggested by Anwar et al. [9], Gabriel et al. [10], and Greshake et al. [29]. In cases where it may not be possible to eliminate the usage of vulnerable generative agents, safeguards must be developed to prevent agents from being actionable when hijacked. For example, sensitive information or services may require secondary human authentication to be accessed, or generative agents may be sandboxed to prevent a hijacked agent from spreading undesirable behaviour to other agents or systems.

Displacement of Human Labour

Discussion of the ethical concerns of any technology would be incomplete without analysis of its effect on human labour. This is especially true for generative agents, as automated generative agents may be highly attractive to organisations that wish to automate tasks that are generally performed by humans. Generative language models have already begun to replace humans in both low- and high-skill occupations, such as customer service agents [35], translators [36], and varying forms of knowledge experts [36, 37]. As a result, generative agents are expected to have significant impact on the quantity and quality of human labour, among other socioeconomic factors [9, 10]

Currently, AI tools are still not sophisticated enough to fully replace humans in many positions [38]. However, the design benefits of generative agents potentially introduce the necessary capabilities to be effectively leveraged in many of these applications. As a result, overzealous usage of generative agents may cause the rapid displacement of human workers across many occupations in a particularly disruptive and unprecedented manner [9]. As such, organisations and researchers should prioritize techniques that use generative agents as collaborative tools to supplement human workers instead of attempting to replace human labour. This approach is validated by the extant literature, which asserts that human-AI collaboration is highly effective at improving productivity while simultaneously not harming the current human workforce [39–43].

Exploitation of Developing Nations and Modern Slavery

As demand grows for generative agents and other GAI-based tools, so will the necessity to manufacture physical hardware and electronic devices capable of leveraging them (such as GPUs and smartphones). Much of the natural and human resources used in the manufacture of these devices comes from developing nations, where there are significant risks of exploitation and contributing to modern slavery [44–47]. Notably, we were unable to find any literature providing meaningful discussion of these concerns as they relate to GAI, despite substantial documentation of the impacts of the physical resources they require. In particular, the mining of silica—a primary component of computer chips—and lithium—a primary component of batteries—and other materials used to manufacture these electronic devices can pose significant environmental and personal health risks for miners and their communities [44, 45, 48–50].

Individuals and organisations wishing to develop and use generative agents or similar tools must ensure that they are not contributing to these risks. This can be done by evaluating the hardware requirements associated with using generative agents and minimising them wherever possible. Future work should aim to improve the efficiency of generative agents and the required hardware to reduce the need for materials and components whose manufacture and usage may contribute to exploitation and modern slavery. Further, independent audits should be conducted on manufacturers and other sections of the supply-chain to ensure adherence to these principles [47]. Finally, requirements for such hardware should be avoided wherever possible, such as through the using simpler and more sustainable techniques or eliminating superfluous usage of generative agents.

Environmental Impact

Alongside increased demand for AI-based tools, the environmental impact and effective carbon footprint of GAI systems have also increased substantially [10, 51–54]. Gabriel et al. [10] identify a range of ways in which the creation and usage of AI assistants, a notable potential application of generative agents, can impact the environment. Specifically, high energy usage during training and inference, emissions embodied by the production of required hardware, and supporting environmentally irresponsible industries and applications all pose substantial environmental risk. Further, the nature and magnitude of embodied emissions associated with model training and the manufacture of required hardware are only just beginning to be investigated and understood [10, 54–57].

To mitigate these concerns, Gabriel et al. [10] suggest a number of emissions-reduction approaches, such as minimising model size, improving hardware efficiency, sourcing carbon-free energy, and implementing public policy to encourage environmentally sustainable development and deployment. However, these approaches heavily rely on the accessibility and feasibility of low-emissions resources and techniques. As such, similar to the discussed concerns about hijacking, we believe a simpler and more feasible approach would be to minimise or eliminate the usage of generative agents wherever possible. Specifically, the implementation of generative agents must be critically analysed with respect to their necessity and weighed against their environmental impact to avoid superfluous or redundant usage [51, 54, 56, 57].

4 Conclusion

In summary, the development and application of generative agents present many ethical challenges and concerns. Ethical problems arise from all sides of GAI usage, including developers, malicious actors, and normal users. Frivolous implementation, unsafe or inefficient design, unreliable and untrustworthy behaviour, and user error all pose significant threats to the ethical usage of generative agents. Given these challenges, continued overzealous acceleration of generative agent development needs to be considered critically and addressed. Currently, the ethical evaluations and impacts of newly developed GAI and generative agent techniques are often left as an afterthought, if they are discussed at all. Thus, future research and applications of generative agents should directly account for the ethical concerns posed in this work and those identified in other works, such as Bail [2], Lazar [7], Chan et al. [8], Anwar et al. [9], and Gabriel et al. [10], among others. To this end, researchers, developers, and legislators should apply the proposed mitigation approaches wherever possible, and create new mitigation techniques where solutions have yet to be developed.

References

[1] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '23, New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1–22, ISBN: 9798400701320. DOI: 10.1145/3586183.3606763. [Online]. Available: https://dl.acm.org/doi/10.1145/3586183.3606763.

[2] C. A. Bail, "Can Generative AI improve social science?" *Proceedings of the National Academy of Sciences*, vol. 121, no. 21, e2314021121, May 2024. DOI: 10.1073/pnas.2314021121. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2314021121.

[3] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," en, *PLOS Medicine*, vol. 2, no. 8, e124, Aug. 2005, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.0020124. [Online]. Available: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124.

[4] P. E. Shrout and J. L. Rodgers, "Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis," en, *Annual Review of Psychology*, vol. 69, no. Volume 69, 2018, pp. 487–510, Jan. 2018, ISSN: 0066-4308, 1545-2085. DOI: 10.1146/annurev-psych-122216-011845. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-psych-122216-011845.

[5] B. J. Wiggins and C. D. Christopherson, "The replication crisis in psychology: An overview for theoretical and philosophical psychology," *Journal of Theoretical and Philosophical Psychology*, vol. 39, no. 4, pp. 202–217, 2019, ISSN: 2151-3341. DOI: 10.1037/teo0000137.

[6] J. L. Tackett, C. M. Brandes, K. M. King, and K. E. Markon, "Psychology's Replication Crisis and Clinical Psychological Science," en, *Annual Review of Clinical Psychology*, vol. 15, no. Volume 15, 2019, pp. 579–604, May 2019, ISSN: 1548-5943, 1548-5951. DOI: 10.1146/annurev-clinpsy-050718-095710. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-clinpsy-050718-095710.

[7] S. Lazar, *Frontier AI Ethics: Anticipating and Evaluating the Societal Impacts of Generative Agents*, en, Apr. 2024. DOI: 10.48550/arXiv.2404.06750. [Online]. Available: https://arxiv.org/abs/2404.06750.

[8] A. Chan et al., "Harms from increasingly agentic algorithmic systems," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23, Chicago, IL, USA: Association for Computing Machinery, 2023, pp. 651–666, ISBN: 9798400701924. DOI: 10.1145/3593013.3594033. [Online]. Available: https://doi.org/10.1145/3593013.3594033.

[9] U. Anwar et al., "Foundational challenges in assuring alignment and safety of large language models," *Transactions on Machine Learning Research*, 2024, ISSN: 2835-8856. [Online]. Available: https://openreview.net/forum?id=oVTkOs8Pka.

[10] I. Gabriel et al., *The ethics of advanced ai assistants*, 2024. arXiv: 2404.16244 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2404.16244.

[11] G. Abercrombie, A. Cercas Curry, T. Dinkar, V. Rieser, and Z. Talat, "Mirages. On Anthropomorphism in Dialogue Systems," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4776–4790. DOI: 10.18653/v1/2023.emnlp-main.290. [Online]. Available: https://aclanthology.org/2023.emnlp-main.290.

[12] A. Choudhury and H. Shamszare, "Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis," *Journal of Medical Internet Research*, vol. 25, no. 1, e47184, Jun. 2023. DOI: 10.2196/47184. [Online]. Available: https://www.jmir.org/2023/1/e47184.

[13] X. Zhan, Y. Xu, and S. Sarkadi, "Deceptive AI Ecosystems: The Case of ChatGPT," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI '23, New York, NY, USA: Association for Computing Machinery, Jul. 2023, pp. 1–6, ISBN: 9798400700149. DOI: 10.1145/3571884.3603754. [Online]. Available: https://dl.acm.org/doi/10.1145/3571884.3603754.

[14] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623, ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. [Online]. Available: https://doi.org/10.1145/3442188.3445922.

[15] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 17 754–17 762, Mar. 2024, ISSN: 2374-3468. DOI: 10.1609/aaai.v38i16.29728. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/29728.

[16] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-Context Retrieval-Augmented Language Models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1316–1331, Nov. 2023, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00605. [Online]. Available: https://doi.org/10.1162/tacl_a_00605.

[17] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24, New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 6491–6501, ISBN: 9798400704901. DOI: 10.1145/3637528.3671470. [Online]. Available: https://dl.acm.org/doi/10.1145/3637528.3671470.

[18] M. Himelein-Wachowiak, S. Giorgi, A. Devoto, M. Rahman, L. Ungar, H. A. Schwartz, D. H. Epstein, L. Leggio, and B. Curtis, "Bots and Misinformation Spread on Social Media: Implications for COVID-19," EN, *Journal of Medical Internet Research*, vol. 23, no. 5, e26933, May 2021. DOI: 10.2196/26933. [Online]. Available: https://www.jmir.org/2021/5/e26933.

[19] N. Hajli, U. Saeed, M. Tajvidi, and F. Shirazi, "Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence," en, *British Journal of Management*, vol. 33, no. 3, pp. 1238–1253, 2022, ISSN: 1467-8551. DOI: 10.1111/1467-8551.12554. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8551.12554.

[20] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," en, *Nature Communications*, vol. 9, no. 1, p. 4787, Nov. 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-018-06930-7. [Online]. Available: https://www.nature.com/articles/s41467-018-06930-7.

[21] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, 7:1–7:41, Jan. 2021, ISSN: 0360-0300. DOI: 10.1145/3425780. [Online]. Available: https://dl.acm.org/doi/10.1145/3425780.

[22] W. M. Lim, "Fact or fake? The search for truth in an infodemic of disinformation, misinformation, and malinformation with deepfake and fake news," EN, *Journal of Strategic Marketing*, Sep. 2023, ISSN: 0965-254X. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/0965254X.2023.2253805.

[23] S. H. Na, S. Cho, and S. Shin, "Evolving Bots: The New Generation of Comment Bots and their Underlying Scam Campaigns in YouTube," in *Proceedings of the 2023 ACM on Internet Measurement Conference*, ser. IMC '23, New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 297–312, ISBN: 9798400703829. DOI: 10.1145/3618257.3624822. [Online]. Available: https://dl.acm.org/doi/10.1145/3618257.3624822.

[24] S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh, "From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models," English, in *2024 IEEE Symposium on Security and Privacy (SP)*, IEEE Computer Society, May 2024, pp. 221–221, ISBN: 9798350331301. DOI: 10.1109/SP54263.2024.00182. [Online]. Available: https://www.computer.org/csdl/proceedings-article/sp/2024/313000a221/1WPcYLpYFHy.

[25] L. Alotaibi, S. Seher, and N. Mohammad, "Cyberattacks Using ChatGPT: Exploring Malicious Content Generation Through Prompt Engineering," in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, Jan. 2024, pp. 1304–1311. DOI: 10.1109/ICETSIS61505.2024.10459698. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10459698.

[26] E. Ferrara, "GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models," en, *Journal of Computational Social Science*, Feb. 2024, ISSN: 2432-2725. DOI: 10.1007/s42001-024-00250-1. [Online]. Available: https://doi.org/10.1007/s42001-024-00250-1.

[27] A. G. Chowdhury, M. M. Islam, V. Kumar, F. H. Shezan, V. Kumar, V. Jain, and A. Chadha, *Breaking Down the Defenses: A Comparative Survey of Attacks on Large Language Models*, Mar. 2024. DOI: 10.48550/arXiv.2403.04786. arXiv: 2403.04786 [cs]. [Online]. Available: http://arxiv.org/abs/2403.04786.

[28] Z. Xu, Y. Liu, G. Deng, Y. Li, and S. Picek, "A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 7432–7449. [Online]. Available: https://aclanthology.org/2024.findings-acl.443.

[29] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, ser. AISec '23, New York, NY, USA: Association for Computing Machinery, Nov. 2023, pp. 79–90, ISBN: 9798400702600. DOI: 10.1145/3605764.3623985. [Online]. Available: https://dl.acm.org/doi/10.1145/3605764.3623985.

[30] I. Kilovaty, "Hacking Generative AI," *Loyola of Los Angeles Law Review*, vol. 58, Mar. 2024. [Online]. Available: https://papers.ssrn.com/abstract=4788909.

[31] P. Levi and C. P. Neumann, "Vocabulary Attack to Hijack Large Language Model Applications," in *CLOUD COMPUTING 2024, The Fifteenth International Conference on Cloud Computing, GRIDs, and Virtualization*, Apr. 2024, pp. 19–24, ISBN: 978-1-68558-156-5. [Online]. Available: https://www.thinkmind.org/library/CLOUD_COMPUTING/CLOUD_COMPUTING_2024/cloud_computing_2024_2_10_28007.html.

[32] K. Notopoulos, "A car dealership added an AI chatbot to its site. Then all hell broke loose.," *Business Insider*, Dec. 2023. [Online]. Available: https://www.businessinsider.com/car-dealership-chevrolet-chatbot-chatgpt-pranks-chevy-2023-12.

[33] L. Debter, "Retailers Are Testing An AI Bot That Haggles With Customers Over Price," *Forbes*, Sep. 2023. [Online]. Available: https://www.forbes.com/sites/laurendebter/2023/09/28/retailers-are-testing-an-ai-bot-that-haggles-with-customers-over-price/.

[34] M. Faithfull, "The Future Of Haggling: Bargain Hunters Negotiate Deals With AI Bot," *Forbes*, Aug. 2024. [Online]. Available: https://www.forbes.com/sites/markfaithfull/2024/08/19/the-future-of-haggling-bargain-hunters-negotiate-deals-with-ai-bot/.

[35] P. Verma, "ChatGPT provided better customer service than his staff. He fired them.," en-US, *Washington Post*, Oct. 2023, ISSN: 0190-8286. [Online]. Available: https://www.washingtonpost.com/technology/2023/10/03/ai-customer-service-jobs/.

[36] E. D. Yilmaz, I. Naumovska, and V. A. Aggarwal, *AI-Driven Labor Substitution: Evidence from Google Translate and ChatGPT*, en, SSRN Scholarly Paper, Rochester, NY, Mar. 2023. DOI: 10.2139/ssrn.4400516. [Online]. Available: https://papers.ssrn.com/abstract=4400516.

[37] D. Kalla, N. Smith, F. Samaah, and S. Kuraku, "Study and Analysis of Chat GPT and its Impact on Different Fields of Study," en, *International Journal of Innovative Science and Research Technology*, vol. 8, no. 3, pp. 827–833, Mar. 2023, ISSN: 2456-2165. [Online]. Available: https://papers.ssrn.com/abstract=4402499.

[38] P. Farhi, "A news site used AI to write articles. It was a journalistic disaster.," en-US, *Washington Post*, Jan. 2023, ISSN: 0190-8286. [Online]. Available: https://www.washingtonpost.com/media/2023/01/17/cnet-ai-articles-journalism-corrections/.

[39] E. Brynjolfsson, D. Li, and L. R. Raymond, *Generative AI at Work*, Working Paper, Apr. 2023. DOI: 10.3386/w31161. [Online]. Available: https://www.nber.org/papers/w31161.

[40] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, *The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems*, en, May 2021. DOI: 10.48550/arXiv.2105.03354. [Online]. Available: https://arxiv.org/abs/2105.03354.

[41] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277–304, Jul. 2023, ISSN: 1522-8053. DOI: 10.1080/15228053.2023.2233814. [Online]. Available: https://doi.org/10.1080/15228053.2023.2233814.

[42] Y. Lai, A. Kankanhalli, and D. Ong, "Human-AI Collaboration in Healthcare: A Review and Research Agenda," *Hawaii International Conference on System Sciences 2021 (HICSS-54)*, Jan. 2021. [Online]. Available: https://aisel.aisnet.org/hicss-54/cl/machines_as_teammates/5.

[43] K. Sowa, A. Przegalinska, and L. Ciechanowski, "Cobots in knowledge work: Human – AI collaboration in managerial professions," *Journal of Business Research*, vol. 125, pp. 135–142, Mar. 2021, ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2020.11.038. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014829632030792X.

[44] J. von der Goltz and P. Barnwal, "Mines: The local wealth and health effects of mineral mining in developing countries," *Journal of Development Economics*, vol. 139, pp. 1–16, Jun. 2019, ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2018.05.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304387818304875.

[45] R. M. Maier, F. Díaz-Barriga, J. A. Field, J. Hopkins, B. Klein, and M. M. Poulton, "Socially Responsible Mining: The Relationship between Mining and Poverty, Human Health and the Environment," *Reviews on environmental health*, vol. 29, no. 0, pp. 83–89, 2014, ISSN: 0048-7554. DOI: 10.1515/reveh-2014-0022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4739650/.

[46] C. Dorninger, A. Hornborg, D. J. Abson, H. von Wehrden, A. Schaffartzik, S. Giljum, J.-O. Engler, R. L. Feller, K. Hubacek, and H. Wieland, "Global patterns of ecologically unequal exchange: Implications for sustainability in the 21st century," *Ecological Economics*, vol. 179, p. 106 824, Jan. 2021, ISSN: 0921-8009. DOI: 10.1016/j.ecolecon.2020.106824. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921800920300938.

[47] C. Han, F. Jia, M. Jiang, and L. Chen, "Modern slavery in supply chains: A systematic literature review," *International Journal of Logistics Research and Applications*, vol. 0, no. 0, pp. 1–22, 2022, ISSN: 1367-5567. DOI: 10.1080/13675567.2022.2118696. [Online]. Available: https://doi.org/10.1080/13675567.2022.2118696.

[48] A. Mishra, "Impact of silica mining on environment," *Journal of Geography and Regional Planning*, vol. 8, no. 6, pp. 150–156, Jun. 2015, ISSN: 2070-1845. DOI: 10.5897/JGRP2015.0495. [Online]. Available: https://academicjournals.org/journal/JGRP/article-abstract/915EC0C53587.

[49] T. Ribeiro, A. Lima, and C. Vasconcelos, "The need for transparent communication in mining: A case study in lithium exploitation," *International Journal of Science Education, Part B*, vol. 11, no. 4, pp. 324–343, Oct. 2021, ISSN: 2154-8455. DOI: 10.1080/21548455.2021.1999530. [Online]. Available: https://doi.org/10.1080/21548455.2021.1999530.

[50] J. Vidal, P. Guest, and g. b. P. Guest, "How developing countries are paying a high price for the global mineral boom," en-GB, *The Observer*, Aug. 2015, ISSN: 0029-7712. [Online]. Available: https://www.theguardian.com/global-development/2015/aug/15/developing-countries-high-price-global-mineral-boom.

[51] P. Dhar, "The carbon impact of artificial intelligence," en, *Nature Machine Intelligence*, vol. 2, no. 8, pp. 423–425, Aug. 2020, ISSN: 2522-5839. DOI: 10.1038/s42256-020-0219-9. [Online]. Available: https://www.nature.com/articles/s42256-020-0219-9.

[52] C.-J. Wu et al., "Sustainable AI: Environmental Implications, Challenges and Opportunities," en, *Proceedings of Machine Learning and Systems*, vol. 4, pp. 795–813, Apr. 2022. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html.

[53] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2023, pp. 1–9. DOI: 10.1109/HPEC58863.2023.10363447. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10363447.

[54] P. Jiang, C. Sonne, W. Li, F. You, and S. You, "Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots," *Engineering*, Apr. 2024, ISSN: 2095-8099. DOI: 10.1016/j.eng.2024.04.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809924002315.

[55] A. Faiz, S. Kaneda, R. Wang, R. C. Osi, P. Sharma, F. Chen, and L. Jiang, "LLMCarbon: Modeling the end-to-end carbon footprint of large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=aIok3ZD9to.

[56] A.-L. Ligozat, J. Lefevre, A. Bugeau, and J. Combaz, *Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions*, Jan. 2022. DOI: 10.3390/su14095172. [Online]. Available: https://www.mdpi.com/2071-1050/14/9/5172.

[57] C. Mulligan and S. Elaluf-Calderwood, "AI ethics: A framework for measuring embodied carbon in AI systems," en, *AI and Ethics*, vol. 2, no. 3, pp. 363–375, Aug. 2022, ISSN: 2730-5961. DOI: 10.1007/s43681-021-00071-2. [Online]. Available: https://doi.org/10.1007/s43681-021-00071-2.