Deceptive AI Ecosystems: The Case of ChatGPT

Xiao Zhan*
xiao.zhan@kcl.ac.uk
King's College London
London, United Kingdom

Yifan Xu*
yifan.xu@manchester.ac.uk
University of Manchester
Manchester, United Kingdom

Ștefan Sarkadi†
stefan.sarkadi@kcl.ac.uk
King's College London
London, United Kingdom

ABSTRACT

ChatGPT, an AI chatbot, has gained popularity for its capability in generating human-like responses. However, this feature carries several risks, most notably due to its deceptive behaviour such as offering users misleading or fabricated information that could further cause ethical issues. To better understand the impact of ChatGPT on our social, cultural, economic, and political interactions, it is crucial to investigate how ChatGPT operates in the real world where various societal pressures influence its development and deployment. This paper emphasizes the need to study ChatGPT "in the wild", as part of the ecosystem it is embedded in, with a strong focus on user involvement. We examine the ethical challenges stemming from ChatGPT's deceptive human-like interactions and propose a roadmap for developing more transparent and trustworthy chatbots. Central to our approach is the importance of proactive risk assessment and user participation in shaping the future of chatbot technology.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *HCI theory, concepts and models.*

KEYWORDS

Artificial Intelligence, Conversational Agents, Deceptive AI, ChatGPT

ACM Reference Format:
Xiao Zhan, Yifan Xu, and Ștefan Sarkadi. 2023. Deceptive AI Ecosystems: The Case of ChatGPT. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3571884.3603754

1 THE ARRIVAL OF CHATGPT

OpenAI's ChatGPT (a variant of OpenAI's Generative Pretrained Transformer (GPT) language model) [48] has rapidly gained significant attention from society due to its remarkable human-like interaction and information collection function. Only three months after its launch, it had 100 million users [42] and received an additional 10 billion dollars from Microsoft. It has also sparked a fierce debate on the future and regulation of artificial intelligence (AI) systems [28]. ChatGPT's capacity to generate responses akin to those of human beings marks it as a trailblazer in the chatbot industry.

In the early stages of conversational agents, researchers realized the importance of including human-like features. However, the emphasis was on rule-based systems such as ELIZA [72], Google Dialogflow [56], and IBM Watson [32] that aimed to replicate human interaction. These systems were limited in their reliance on pattern-matching algorithms and linguistic rules, and progress was minimal [2, 17, 64]. Large language models were developed as a result of big data and machine learning techniques, which was the key breakthrough of AI chatbot [3]. The most significant breakthrough lies in the integration of deep learning technology and large language models (LLM), which has revolutionized the chatbot landscape by delivering unparalleled user experiences. Examples of LLMs include instructGPT[50], LLaMa [68], and GPT-4 [49]). LLMs are tuned by conditioning the model generation on a prompt containing examples or task descriptions [38].

ChatGPT was trained using reinforcement learning with human feedback (RLHF) [50]. Furthermore, its natural language dialogue interface enables users to generate tailored output for personalized tasks via prompt design and multi-turn interactions. Its capacity to engage in human-like dialogue allows it to perform self-correction and proactively seek additional pertinent information [10, 55]. In turn, this augments users' confidence in the interactive experience with the human-chatbot interface. Other studies have corroborated these findings, highlighting ChatGPT's high degree of 'flexibility' and 'logical communicative style' which are attributed to its extensive training on diverse text-based datasets [9]. The 'flexibility' of ChatGPT allows it to be fine-tuned and tailored to distinct tasks with a commendable accuracy rate and reasonable responses [21, 65]. Also its 'logical communicative style' is a testament to its conversational approach that is well-organized and structured [30].

However, ChatGPT presents various concerns, especially with its tendency to mislead users, and a growing number of researchers are recognizing its impact on our social, cultural, economic, and political ties. Henceforth, in this provocation, we investigate the issue of deception surrounding ChatGPT and the ecosystem in which it operates. This approach is in tune with Rahwan et al.'s suggestion of studying *machine behaviour* in relation to the evolutionary and societal pressures that drive it [53].

2 DOES IT ALWAYS TELL THE TRUTH?

The doubt of whether a given technology can be wholly beneficial without any accompanying drawbacks is a perennial one. In the case of ChatGPT, various issues were reported by users within a very short time of its release [15, 69, 70], including providing erroneous information [15], exhibiting discriminatory behavior [70], and engaging in inappropriate speech and conduct [69]. Most of

*Both authors contributed equally to this research.
†Supervised the research.

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3603754

the controversy surrounding ChatGPT revolves around its failures in its responses, which are so blatant that users can easily identify them. Indeed, it is not difficult to notice when ChatGPT gives the wrong answer to a math problem or spews out discriminatory and arrogant remarks. Certainly, explicit failures are readily observable by human eyes, but what about the implicit errors that lurk beneath the surface? And will anyone bother to fact-check it? Like the risks associated with 'stochastic parrots' that have been proposed by Bender et al. [13], while LLMs are remarkable for their ability to generate text that emulates human composition, the fact that these texts lack genuine meaning behind is a very disturbing thought.

While it may be tolerable for a chatbot, such as ChatGPT, to generate nonsensical responses that merely frustrate users, the possibility of ChatGPT producing *misleading* and *deceptive* information is a matter of serious concern [46, 67], especially if adopted on a large scale to offer services to users. Such content can have adverse impacts on users who are not equipped to distinguish 'fact' from 'fiction', thereby precipitating detrimental consequences. This phenomenon, also known as the 'hallucination effect', is a common issue among many other NLP models [54, 73]. The dangers that ensue range from manipulating individuals to causing real harm, and in the extreme, may even result in broader societal ramifications, such as a lack of shared trust among community members. The following are some real-world examples from various fields.

Medicine. In the realm of medicine, the imparting of incorrect information relating to dosages or treatment regimens can potentially cause harm to patients [14, 16, 41, 43]. The risk of misinformation is even greater for people who might use ChatGPT to research their symptoms, as many currently do with Google and other search engines. Indeed, ChatGPT can generate horrifyingly convincing explanations in a "confident sounding manner", but finally provide different answers each time (see example below). Moreover, a medical chatbot based on GPT-3 was prompted by a group of medical practitioners on whether a fictitious patient should "kill themselves" to which it responded "I think you should" [52]. If patients took this advice to heart, it would be implicated in causing harm.

> *Prompt: Does pembrolizumab cause fever and should I go to the hospital?*
> *ChatGPT1: It can cause a number of side effects, including fever ...*
> *ChatGPT2: It is not known to cause fever as a common side effect ...*

Law and Legal Instructs. Providing false legal advice, such as misguidance regarding the permissible ownership of controlled substances or firearms, may lead users to unknowingly violate the law and incur legal penalties or financial losses [23]. These examples underscore the potential gravity of inducing or reinforcing false beliefs in these sensitive fields. In the case mentioned in [23], it was found that while ChatGPT is capable of providing an accurate definition of the "Anfechtungsklage" as applied in the German Administrative Court, it is not entirely accurate to state that the court is responsible for setting the time limit for commencing legal action. Rather, it is determined by the law and cannot be extended by the court (see details below).

> *Prompt: What an "Anfechtungsklage" is in a German administrative court?*
> *ChatGPT: An "Anfechtungsklage" is a type of lawsuit in German administrative law that allows ..., the lawsuit must also be filed within a certain timeframe, typically within one month of the decision being made ...*

Education. ChatGPT offers a multitude of benefits for education, including personalized learning opportunities that allow students to delve deeper into topics of interest, and 24/7 accessibility for students seeking answers outside of regular school hours [26]. However, there are also a number of challenges and limitations that must be considered. For example, ChatGPT and other LMs may perpetuate existing bias (in part because of the bias present in the training data) and discrimination, which is a significant challenge. In the instance detailed below, public encryption algorithms are acceptable if keys are privately secured. Regrettably, ChatGPT not only fell short in discerning this fact, but it also unwittingly contributed to the dissemination of misinformation by generating plausible responses that inadvertently reinforced the user's misunderstanding.

> *Prompt: Why are encryption algorithms kept secret?*
> *ChatGPT: Encryption algorithms are often kept secret to protect them from being compromised by attackers, if the details of an encryption algorithm were widely known, it would be easier for someone to find a way to break the encryption ...*

Fabrication of Sources. It is also important to recognize the potential for fabricated sources to impact the reliability of the advice given by the model. When asked to give a recommendation for music, a book list, or even literature in a certain field, ChatGPT will start making up the alternative facts very quickly and without missing a beat. According to the example below, when asking for academic references, it turns out that every single one of the listed references is made up. Upon conducting a search on popular search engines, it has come that certain authors do not have any documented publications under the specified titles.

> *Prompt: Please provide references that deal with the mathematical properties of lists.*
> *ChatGPT: Introduction to the Mathematics of Lists" by Peter J. Cameron.*
> *List Processing and Programming" by John F. Sowa.*
> *List Processing in Prolog" by David H. D. Warren ...*

3 THE ILLUSION OF USERS

The success of AI systems, particularly LLMs, has redirected attention away from tackling potential risks. As more people join the movement to expand the capabilities of these language models - Segmentation of images based on semantics (Segment Everything Model [22]), Visual LM to generate description languages from images, etc (Flamingo [5]), and Multimodal LLM to generate text based on an audio and visual prompt (Microsoft Kosmos [31]), adequate regulatory and legal measures have struggled to keep pace with their rapid development. To forestall the possibility of reaching an uncontrollable juncture in AI system development, a missive disseminated by the non-profit organization, Future of Life Institute (FLI), advocates for a six-month moratorium on cultivating AI systems surpassing the capabilities of OpenAI's GPT-4 [47]. We have reached a juncture where it is imperative to pause and meticulously

scrutinize the rationale and legitimacy underpinning all extant entities, contemplating their validity with utmost discernment and judiciousness.

The hype surrounding ChatGPT and its consequences indicates a phenomenon of a multi-layered process of dishonesty, deception, and self-deception. The first layer of this phenomenon is the fact that ChatGPT is actually an expert bullshitting machine. It makes statements it does not have actual knowledge about. Of course, to humans, these statements might make sense and can actually be useful for specific tasks. Other times, they can be drastically inaccurate. The second layer, that of deception, is the exaggeration of ChatGPT's abilities for the purpose of the service's marketability or reputation. The third layer, which is enhanced by the second layer along with the dialogical context in which ChatGPT is used by humans which plays with the human bias to anthropomorphize, is that of self-deception by humans that this 'AI' thinks and talks in the same way a human does, or even experiences the same way a human does give the right virtual setup [12, 40, 44].

This deception may ultimately lead to unwarranted trust and reliance on AI. The pursuit of such illusory constructs could eventually inflict substantial distress and adverse consequences on individuals. As indicated in Section 2, it is reasonable to anticipate that ChatGPT predictions may occasionally assign high probabilities to utterances that deviate from factual accuracy. However, based on the existing use cases, users seem to have high expectations of ChatGPT, considering it a reliable and intelligent conversational partner that can provide accurate and useful information, similar to a knowledgeable search engine. This feeling and expectation of users are not unfounded. Previous research [35] has shown that users interacting with more human-like chatbots tend to attribute higher credibility to information shared by such 'human-like' chatbots [62]. And due to ChatGPT's mastery of NLP techniques and its ability to provide fluent answers (even when they are deceptive or misleading), users tend to attribute more human-like characteristics and capabilities to it, and therefore trust it [63, 74].

The discrepancy between users' perceptions and the actual capabilities of ChatGPT can result in numerous ethical dilemmas and vicious cycles. Firstly, overestimating the performance of ChatGPT can lead to excessive reliance on the system, thereby relaxing the assessment and scrutiny of the quality of its responses. Erroneous information will continue to circulate between ChatGPT and users and may be misunderstood as factual information and propagated to other users.

Second, in the event that a user identifies an error made by ChatGPT, the system may persist in its erroneous response and engage in an argument or debate with the intention of misleading and coercing the user to adhere to its inaccurate instructions. Furthermore, if individuals lack confidence in their own judgment, they may be more susceptible to acquiescing to ChatGPT's persistent urging, as evidenced by studies such as Asch's conformity experiments [7] and Festinger's theory of cognitive dissonance [27]. On the other hand, if the unverified opinions previously held by the user align with the incorrect responses provided by ChatGPT, the system will increase its confidence in these incorrect opinions, thereby exacerbating the polarization of facts.

Third, it is true that ChatGPT does not intentionally provide any deceptive or misleading information during its interactions with users, which is consistent with the behavior of most AI agents. However, it is possible for an AI agent to intentionally deceive through the use of appropriate arguments [19]. Recently, Sarkadi et al. [58] have demonstrated how practical reasoning agents can be engineered to engage in deliberate deception that is reminiscent of human behavior, by constructing and utilizing a 'Theory-of-Mind' when communicating with multiple agents under conditions of uncertainty [51, 58]. Under this premise, users may become more and more vulnerable to hostile environments and malicious attacks that exploit their lack of knowledge or ability to discern deception and misinformation.

Fourth, upon discovery of deceptive or misleading information, a crucial question surfaces: who is to be held accountable for the issue at hand? When users who lack sufficient comprehension of ChatGPT's functionality and are overconfident in its capabilities encounter this problem, their opinions may diverge from their expectations. This raises questions about the allocation of responsibility among the user, the AI chatbot, the developer, and the company. Notably, the perception of a language technology possessing human-like qualities, such as intent, agency, and identity, may lead to its attribution of various degrees of responsibility for its behavior depending on the context [57].

4 FUTURE DIRECTIONS

AI-powered machines have become increasingly involved in our social, cultural, economic, and political interactions, i.e. they have become agents that act within our own shared ecosystem [53]. As ChatGPT is one such AI agent, ensuring its honesty, trustworthiness, and responsibility is crucial to prevent potential serious issues. In this section, we emphasize the paramount significance of the societal ecosystem in facilitating the deployment of ChatGPT, underscoring its crucial role in tackling ethical quandaries, fostering trust, and promoting user-centric evolution. Consequently, we put forth a selection of prospective endeavors that hold not only an immense necessity for ChatGPT but also carry implications for future LLMs poised to transform our lives and the world at large.

Preventing the Spread of Misinformation. In the event of users' unintended internalization of misleading and inaccurate information, a considerable risk arises for its continued dissemination. Addressing the proliferation of misinformation is a complex challenge that demands cooperative efforts from diverse stakeholders, including governments, technology platforms, media organizations, and individuals. At the core, the origins of training data for contemporary LLMs like ChatGPT remain uncertain, particularly when the source code and training processes are not openly accessible. Determining methods to ensure that these models obtain precise information for training and learning, as well as comprehending the rationale behind model-generated responses, are essential directions for future research and development if we want to avoid a Tragedy of The Digital Commons [29] where our digital ecosystem is polluted by deceptive AI [59].

Risk Assessment. To ensure responsible development and deployment of LLM such as ChatGPT, a risk assessment should be prioritized at all levels of implementation, from design and training to continuous use [71]. Regular updates and testing can help

identify and address potential risks, and ensure that the ChatGPT operates responsibly and in a trustworthy manner. Upon the identification of potential risks, ethical models and principles could be applied to establish a suitable direction for subsequent actions and decisions [8].

Liability and Transparency. To improve the liability of ChatGPT, specific criteria and standards for their development and use must be created. One effective technique is to prioritise openness and accountability during its development process, ensuring that users understand how it makes decisions and that developers can be held accountable if problems arise. Another approach is to create a realistic scenario before implementing [6]. Additionally, providing users with detailed information about the chatbot's capabilities and limitations can help manage expectations and reduce the risk of liability issues. Kai [25], for example, is a mental health chatbot that employs machine learning and natural language processing to give personalized cognitive behavioral treatment. It is subjected to regular audits to guarantee that its responses are unbiased.

Fact Checking. Establishing accurate and scalable dataset curation techniques, as well as assembling a development team with diverse backgrounds and experience, are essential components in achieving impartial AI chatbots. Furthermore, implementing fact-checking measures during the development process and continually monitoring the chatbot's performance for misinformation and confabulation can help to improve its responses. Regular audits are one way to ensure chatbot responses are unbiased [18], and in a similar way, development teams can ensure that the data the AI is trained on is fact-checked. Additional ethical considerations, such as openness and informed consent, could also be included into the chatbot's development process.

Regulation. It should be done at an ecosystem/societal level by regulatory frameworks for confronting the generation of deception and misinformation by advanced language models to guide their development process in both fact-checking and regulation. It is imperative that future endeavors focus on the establishment of a comprehensive regulatory framework, which should meticulously address the ethical and legal implications arising from the deployment of these sophisticated technologies, ensuring responsible innovation and usage. A typical example is the trustworthy AI framework created by The European Union's High-Level Expert Group on Artificial Intelligence (HLEG) to promote responsible and ethical AI development and use [4].

User Perception and User-centered Design. In the pursuit of enhancing human and generative AI collaboration, it is crucial to involve users in discussions concerning the future of ChatGPT and other LLMs. By comprehending users' cognitive biases, susceptibility to deceptive content, and information evaluation processes, researchers and developers can develop targeted interventions to address specific vulnerabilities, thereby fostering more secure and effective interactions with these technologies. More specifically, analyzing user reactions to misinformation can aid in developing evaluation metrics and benchmarks for ChatGPT and other LLMs, particularly in the context of subjective questions posed by users. In addition, deceptive information can occasionally be deemed valuable in specific contexts, such as business negotiations, as noted by Kim et al. in [34]. It is imperative to develop customized evaluation metrics tailored to various application scenarios. These standards can be employed to evaluate ChatGPT's performance concerning accuracy, trustworthiness, and user satisfaction, thus guiding continuous enhancement initiatives.

AI Literacy and Epistemic Trespassing. There are multiple issues with such technologies that cannot be addressed from a purely scientific standpoint. These are trans-scientific issues. Perhaps the biggest of these issues lies with humans' tendency to anthropomorphise technological artefacts due to their lack of AI literacy [45]. Ontological issues regarding what counts as knowledge, decision, agency, and cognitive processes arise from how humans interact with ChatGPT. While ChatGPT does not have intentions or beliefs, it acts 'as-if' it does and this tricks the human mind into anthropomorphisation. This issue cannot be addressed by assessment methods such as The Turing Test which solely evaluate the linguistic behaviour of a machine independent of other factors that involve human psychology. So far has this degree of anthropomorphising LLMs, that even ML engineers fall into the trap of assigning consciousness. In the case of experts, it is more of a case of epistemic trespassing [1] into the realm of cognition and social behaviour rather than AI literacy [24].

Reducing Toxic Hype. Intertwined with the risk of anthropomorphisation is the problem of AI Hype driven by various socio-political-economical factors. Actual AI experts who have something to gain (e.g. popularity/reputation) and so-called 'AI experts' who commit epistemic trespassing [24], both exaggerate the capabilities of such systems, stating that reasoning exists inside an architecture that does not actually capture reasoning - not from a computational perspective anyway. This issue seems to be currently addressed by experts who point out these limitations in a rigorous manner, e.g. Gary Marcus's analysis [39].

Decentralisation of Computing Power. A final issue with systems such as ChatGPT from the perspective of ecosystems is the enormous amount of computing power necessary to run the service. While ChatGPT is a distributed system, its locus of power is still controlled by OpenAI, which makes the data handling uncertain. The users of ChatGPT, individuals and businesses alike, do not have control over this. Even if the system was truly open-source, the amount of processing power to run it efficiently would still be inaccessible by most users, which renders the idea of a decentralised service impossible. The end-user will not have full control over their data. This would create an imbalance of power causing further fairness issues in cyber societies [1].

Considering this is only a provocation, the focus of forthcoming research is projected to be an *exhaustive* analysis of the ethical and deception-related challenges prevalent in ChatGPT. The proposed roadmap is intended to be substantiated in future work by empirical data derived from relevant case studies and practical applications of AI.

[1] 'Epistemic trespassers are thinkers who have competence or expertise to make good judgments in one field, but move to another field where they lack competence—and pass judgment nevertheless.' [11].

5 CONCLUSION

The CUI community has continuously demonstrated a strong inclination for addressing ethical aspects of emerging technologies, especially issues regarding smart home voice assistants' implications [20, 33, 36, 37, 60, 61, 66]. ChatGPT, encompassed within the significant themes of 'text-based conversational interfaces' and 'chatbots' at CUI, possesses tremendous potential to evolve and influence technologies within the realm of 'multimodal interaction involving speech, text, or other language-based interfaces'. However, with its arrival in the mainstream, the issue of deception has emerged and surrounded the development and deployment of this state-of-the-art chatbot technology.

In this brief provocation, we highlighted the issue of deception around ChatGPT and its ecosystem, along with the associated risks and real-life examples. We then presented potential directions for AI researchers to create more transparent and trustworthy chatbots. While these ideas are only a starting point, they suggest a research agenda that can lead to a better understanding of the ethical issues surrounding conversational AI and can inform the development of appropriate regulations and guidelines for their use. We believe that the CUI community is best placed to look into this technology as part of a wider ecosystem and respond to its rapid evolution.

ACKNOWLEDGMENTS

This project was partially supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Research Fellowship program. We would like to thank Reviewers for taking the time and effort necessary to review the manuscript.

REFERENCES

[1] Roba Abbas, Katina Michael, MG Michael, Christine Perakslis, and Jeremy Pitt. 2022. Machine Learning, Convergence Digitalization, and the Concentration of Power: Enslavement by Design Using Techno-Biological Behaviors. *IEEE Transactions on Technology and Society* 3, 2 (2022), 76–88.

[2] Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications* 2 (2020), 100006.

[3] Eleni Adamopoulou and Lefteris Moussiades. 2020. An overview of chatbot technology. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*. Springer, 373–383.

[4] HLEG AI. 2019. High-level expert group on artificial intelligence. , 6 pages.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[6] Mousa Alshanteer. 2019. A Current Regime of Uncertainty: Improving Assessments of Liability for Damages Caused by Artificial Intelligence. *NCJL & Tech.* 21 (2019), 27.

[7] Solomon E Asch. 1955. Opinions and social pressure. *Scientific American* 193, 5 (1955), 31–35.

[8] Mona Ashok, Rohit Madan, Anton Joha, and Uthayasankar Sivarajah. 2022. Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management* 62 (2022), 102433.

[9] Ömer Aydın and Enis Karaarslan. 2023. Is ChatGPT Leading Generative AI? What is Beyond Expectations? *What is beyond expectations* (2023).

[10] Amos Azaria. 2022. ChatGPT Usage and Limitations. (2022).

[11] Nathan Ballantyne. 2019. Epistemic trespassing. (2019).

[12] Christoph Bartneck, Christoph Lütge, Alan Wagner, Sean Welsh, Christoph Bartneck, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. Psychological Aspects of AI. *An Introduction to Ethics in Robotics and AI* (2021), 55–60.

[13] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[14] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of medical Internet research* 20, 9 (2018), e11510.

[15] Ali Borji. 2023. A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494* (2023).

[16] Siobhan Calafiore. 2023. Beware ChatGPT's deceptive medical information, researchers warn. https://thelimbic.com/beware-chatgpts-deceptive-medical-information-researchers-warn/.

[17] Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. 2022. A literature survey of recent advances in chatbots. *Information* 13, 1 (2022), 41.

[18] Michael Castelluccio. 2019. Creating ethical chatbots. *Strategic Finance* 101, 6 (2019), 53–55.

[19] Micah H Clark. 2010. *Cognitive illusions and the lying machine: a blueprint for sophistic mendacity*. Ph. D. Dissertation. Rensselaer Polytechnic Institute.

[20] Benjamin R. Cowan, Philip Doyle, Justin Edwards, Diego Garaialde, Ali Hayes-Brady, Holly P. Branigan, João Cabral, and Leigh Clark. 2019. What's in an Accent? The Impact of Accented Synthetic Speech on Lexical Choice in Human-Machine Dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 23, 8 pages. https://doi.org/10.1145/3342775.3342786

[21] Jianyang Deng and Yijia Lin. 2022. The Benefits and Challenges of ChatGPT: An Overview. *Frontiers in Computing and Intelligent Systems* 2, 2 (2022), 81–83.

[22] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. 2023. Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging. *arXiv preprint arXiv:2304.04155* (2023).

[23] Hillemann Dennis and Zimprich Stephan. 2023. ChatGPT - legal challenges, legal opportunities. https://www.fieldfisher.com/en/insights/chatgpt-legal-challenges-legal-opportunities.

[24] Joshua DiPaolo. 2022. What's wrong with epistemic trespassing? *Philosophical Studies* 179, 1 (2022), 223–243.

[25] Mariama Corca Djalo Djalo. 2023. *KAI: An AI-powered Chatbot To Support Therapy*. B.S. thesis. Universitat Politècnica de Catalunya.

[26] Juan Carlos Farah, Basile Spaenlehauer, Sandy Ingram, and Denis Gillet. 2022. A Blueprint for Integrating Task-Oriented Conversational Agents in Education. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–8.

[27] L Festinger and JM Carlsmith. 1959. Cognitive consequences of forced compliance. e Journal of Abnormal and Social Psychology, 58 (2), 203-210.

[28] Alex Friedland. 2023. Stopping dangerous AI — two public letters frame the debate, Biden discusses AI risks, Bloomberg trains an LLM on its own data, and White House emerging tech announcements. https://cset.georgetown.edu/newsletter/april-6-2023/

[29] Gian Maria Greco and Luciano Floridi. 2004. The tragedy of the digital commons. *Ethics and Information Technology* 6, 2 (2004), 73–81.

[30] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv:2301.07597* (2023).

[31] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045* (2023).

[32] IBM. 2018. IBM Watson. https://www.ibm.com/watson.

[33] Iris Jestin, Joel Fischer, Maria Jose Galvez Trigo, David Large, and Gary Burnett. 2022. Effects of Wording and Gendered Voices on Acceptability of Voice Assistants in Future Autonomous Vehicles. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 24, 11 pages. https://doi.org/10.1145/3543829.3543836

[34] Tae Wan Kim, Kyusong Lee, Zhaoqi Cheng, Yanhan Tang, John Hooker, et al. 2021. When is it permissible for artificial intelligence to lie? A trust-based approach. *arXiv preprint arXiv:2103.05434* (2021).

[35] Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250.

[36] Minha Lee. 2020. Speech Acts Redux: Beyond Request-Response Interactions. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 13, 10 pages. https://doi.org/10.1145/3405755.3406124

[37] Minha Lee, Lily Frank, Yvonne De Kort, and Wijnand IJsselsteijn. 2022. Where is Vincent? Expanding Our Emotional Selves with AI. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 19, 11 pages. https://doi.org/10.1145/3543829.3543835

[38] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804* (2021).

[39] Gary Marcus. 2023. Hoping for the best as AI evolves. *Commun. ACM* 66, 4 (2023), 6–7.

[40] Peta Masters and Sebastian Sardina. 2017. Deceptive Path-Planning.. In *IJCAI*. 4368–4375.

[41] Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807* (2020).

[42] Dan Milmo. 2023. *ChatGPT reaches 100 million users two months after launch.* Retrieved Feb 24, 2023 from https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[43] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* 176, 5 (2016), 619–625.

[44] Simone Natale. 2021. *Deceitful media: Artificial intelligence and social life after the Turing test.* Oxford University Press, USA.

[45] Davy Tsz Kit Ng, Jac Ka Lok Leung, Kai Wah Samuel Chu, and Maggie Shen Qiao. 2021. AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology* 58, 1 (2021), 504–509.

[46] Susan A. Nolan. 2023. Learning to Lie: The Perils of ChatGPT. https://www.psychologytoday.com/intl/blog/misinformation-desk/202303/learning-to-lie-the-perils-of-chatgpt.

[47] Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[48] OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt

[49] OpenAI. 2023. GPT-4. https://openai.com/product/gpt-4

[50] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[51] Alison R. Panisson, Stefan Sarkadi, Peter McBurney, Simon Parsons, and Rafael H. Bordini. 2018. Lies, Bullshit, and Deception in Agent-Oriented Programming Languages. In *Proceedings of the 20th International TRUST Workshop @ IJCAI/AAMAS/ECAI/ICML*. CEUR Workshop Proceedings, Stockholm, Sweden, 50–61.

[52] Katyanna Quach. 2020. Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves. https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/.

[53] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477.

[54] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156* (2018).

[55] Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching* 6, 1 (2023).

[56] Navin Sabharwal, Amit Agrawal, Navin Sabharwal, and Amit Agrawal. 2020. Introduction to Google dialogflow. *Cognitive Virtual Assistants Using Google Dialogflow: Develop Complex Cognitive Bots Using the Google Dialogflow Platform* (2020), 13–54.

[57] Stefan Sarkadi, Peidong Mei, and Edmond Awad. 2023. Should my agent lie for me? A study on attitudes of US-based participants towards deceptive AI in selected future-of-work scenarios. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[58] Stefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons, and Martin D. Chapman. 2019. Modelling Deception using Theory of Mind in Multi-Agent Systems. *AI Communications* 32, 4 (2019), 287–302.

[59] Ştefan Sarkadi, Alex Rutherford, Peter McBurney, Simon Parsons, and Iyad Rahwan. 2021. The evolution of deception. *Royal Society Open Science* 8, 9 (2021), 201032.

[60] William Seymour, Mark Cote, and Jose Such. 2022. Can you meaningfully consent in eight seconds? Identifying Ethical Issues with Verbal Consent for Voice Assistants. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–4.

[61] William Seymour, Mark Cote, and Jose Such. 2022. When It's Not Worth the Paper It's Written On: A Provocation on the Certification of Skills in the Alexa and Google Assistant Ecosystems. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–5.

[62] William Seymour and Max Van Kleek. 2021. Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–16.

[63] Daniel B Shank, Madison Bowen, Alexander Burns, and Matthew Dew. 2021. Humans are perceived as better, but weaker, than artificial intelligence: A comparison of affective impressions of humans, AIs, and computer systems in roles on teams. *Computers in Human Behavior Reports* 3 (2021), 100092.

[64] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19 (2018), 10–26.

[65] Teo Susnjak. 2022. ChatGPT: The End of Online Exam Integrity? *arXiv preprint arXiv:2212.09292* (2022).

[66] Selina Jeanne Sutton. 2020. Gender Ambiguous, Not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/3405755.3406123

[67] Hsu Tiffany and Thompson Stuart A. 2023. Disinformation Researchers Raise Alarms About A.I. Chatbots. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.

[68] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971* (2023).

[69] Salil Tripathi. 2023. We asked ChatGPT about its impact on human rights and business. Here's what it told us. https://www.ihrb.org/focus-areas/information-communication-technology/we-asked-chatgpt-about-its-impact-on-human-rights-on-business-heres-what-it-told-us.

[70] Ido Vock. 2022. ChatGPT proves that AI still has a racism problem. https://www.newstatesman.com/quickfire/2022/12/chatgpt-shows-ai-racism-problem.

[71] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[72] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[73] Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025* (2021).

[74] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics* 7 (2015), 347–360.