On Moral Manifestations in Large Language Models

JOEL WESTER, Human Centered Computing Group, Aalborg University, Denmark
JULIEN DELAUNAY, Inria/IRISA, Rennes University, France
SANDER DE JONG, Human Centered Computing Group, Aalborg University, Denmark
NIELS VAN BERKEL, Human Centered Computing Group, Aalborg University, Denmark

Since OpenAI released ChatGPT, researchers, policy-makers, and laypersons have raised concerns regarding its false and incorrect statements, which are furthermore expressed in an overly confident manner. We identify this flaw as part of its functionality and describe why large language models (LLMs), such as ChatGPT, should be understood as social agents manifesting morality. This manifestation happens as a consequence of human-like natural language capabilities, giving rise to humans interpreting the LLMs as potentially having moral intentions and abilities to act upon those intentions. We outline why appropriate communication between people and ChatGPT relies on moral manifestations by exemplifying 'overly confident' communication of knowledge. Moreover, we put forward future research directions of fully autonomous and semi-functional systems, such as ChatGPT, by calling attention to how engineers, developers, and designers can facilitate end-users sense-making of LLMs by increasing moral transparency.

Additional Key Words and Phrases: ChatGPT, large language model, social agent, moral, manifest, moral cognition, overconfident

ACM Reference Format:
Joel Wester, Julien Delaunay, Sander de Jong, and Niels van Berkel. 2018. On Moral Manifestations in Large Language Models. In *CHI '23: ACM CHI Conference on Human Factors in Computing Systems @ Workshop on Moral Agents, April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Since the public release of ChatGPT and other large language models (LLMs), the media has been quick to point to its potential impact on various domains (e.g., student reports, scientific articles). Moreover, concerns have been raised regarding the generation of false or wrongful knowledge. For example, ChatGPT has generated fake scientific abstracts that were able to deceive scientists into believing these abstracts were written by their peers [2]. Despite its ability to produce text in a human-like style, the answers provided can be wrong or based on non-existing sources [12].

Nonetheless, LLMs, such as ChatGPT, communicates in human-like ways, causing humans to anthropomorphise the system [9]. This gives rise to new perceptions of the system, in which constructs traditionally used to describe humans are now applied to systems (e.g., the belief that the other entity has a sense of what is right or wrong). While anthropomorphism has been a popular topic within HCI and related fields, the moral dimensions of anthropomorphism are not well understood. Fiske et al., point out that human perception of interpersonal behaviour helps to determine other people's intentions and their capacity to realise these intentions [4]. These intentions (and the capacity to realise them) have moral dimensions, meaning that intentions (and the capacity to realise them) can be perceived and categorised as right or wrong. Consequently, this may influence human sense-making in ways we do not yet understand. We suggest that this understanding is particularly relevant in people's interactions with LLMs, given their human-like communicative abilities and the growing application of LLMs in sensitive settings such as mental health [11].

The use of highly human-like language (e.g., overconfident communication style) by autonomous systems influences end-user perceptions. Various researchers have raised concerns about ChatGPT's overconfidence in its answers to user prompts. The style of these answers has been described by people using negatively loaded metaphorical language such as 'your average tech-dude' or 'over-confident dudes with shallow subject matter knowledge but high skill in faking expertise' [3, 8]. This illustrates that humans might perceive LLMs as overly-confident individuals, resonating with the quality of interaction and potentially influencing it negatively and giving rise to discontinuing the interaction (e.g., conversational break-down [10]). Moreover, overconfident answers may influence interactions in other ways we cannot yet anticipate, consequently posing risks to humans engaging with LLMs available to the public across a variety of tasks (e.g., asking for advice or writing assistance). We highlight that possible consequences are not only ethical in nature but also pose moral concerns.

2 LARGE LANGUAGE MODELS AS MORAL MANIFESTORS

As outlined above, people tend to anthropomorphise systems (e.g., LLMs) displaying natural language capabilities. When people anthropomorphise these systems, it naturally follows that people use their sense-making abilities to understand their experiences, as they would in making sense of other people around them. Therefore, we mean that communicative autonomous systems, such as large language models, belong to the space of social agents, in which both humans and non-humans are included in terms of functionality. As a consequence, LLMs should be understood as moral manifestors, manifesting moral dimensions through natural language at the same level as humans:

> "By way of analogy, consider the concept VEHICLE. At a mechanical level, vehicles are extremely variable and not at all distinct from other things. A motorcycle, for example, has more in common with a lawn mower than with a sailboat, and a sailboat has more in common with a kite than with a motorcycle. One might conclude from this that the concept VEHICLE is therefore meaningless, but that would be mistaken. VEHICLES are bound together, not at the mechanical level, but at the functional level. I believe that the same is true of morality" [6, p. 40]

As such, seeing LLMs as members of this shared space populated by social agents, we should anticipate that humans interacting with such communicative autonomous systems use the same cognitive functions as when making sense of other humans (relying on cognitive processes such as perception, reasoning, or information processing). Moreover, this implies that humans also have expectations, attitudes, or perspectives that influence how they make sense of systems in different contexts.

How the anthropomorphising of publicly available LLMs (e.g., perceived overconfidence in ChatGPT's answers) affects user interaction is still unclear. Therefore, the moral dimensions of overly confident responses, particularly when they are false or inappropriate, may have a discernible impact on human sense-making processes (e.g., an individual may have reduced trust in a system that provides incorrect responses [13]). Without a doubt, this has implications for a number of stakeholders, including end-users, engineers, and designers. We are breaking new ground in that lines between human and AI written text are getting blurry. This is illustrated by OpenAI, which recently described its new AI classifier to detect AI-written text (text primarily generated by ChatGPT, also developed by OpenAI):

> “Our classifier is not fully reliable. In our evaluations on a ‘challenge set’ of English texts, our classifier correctly identifies 26% of AI-written text (true positives) as ‘likely AI-written,’ while incorrectly labeling human-written text as AI-written 9% of the time (false positives).”

Together with other researchers, we highlight that this ‘breaking new ground’ needs to happen responsibly. The human-centred approach to computing aligns well with having the user in focus and is therefore a suitable approach to follow up on suggested implications. The HCI community should therefore seek to comprehend how LLMs shape, influence, or even harm their users through the use of fully autonomous and semi-functional communicative autonomous systems such as ChatGPT. To do so requires us to see ChatGPT and similar systems as social agents manifesting morality that can be designed and manipulated. More specifically, we suggest that moral manifestations should be understood through the psychological and cognitive processes surrounding morality [6]. This requires us to presume communicative autonomous systems as moral manifestors similar to how we pose humans as moral manifestors.

3 FUTURE RESEARCH DIRECTIONS

In general, deploying large language models into the wild results in a wild variety of conversational topics. By considering large language models as social agents manifesting morality, researchers can better assess their impact on end-user perceptions and provide concrete recommendations for designing these systems across various contexts. Future research must seek to align LLMs moral manifestations through social communication abilities with end-user expectations (e.g., by avoiding overconfidence in its communication). Currently, LLMs manifest morality by providing (overconfident communication, potentially false or wrongful) answers we do not want in ways we do not like. We suggest that by understanding LLMs as moral manifestors and by introducing moral transparency, chances are increased of meeting end-user expectations.

More specifically, we call attention to the need to increase moral transparency (i.e., morally relevant transparency [7]) of LLMs. Increasing moral transparency can partly be achieved by referencing the answers provided. Similarly to the argument put forward by Glaese et al., this allows users to assess the source of the answer and to form an opinion about its correctness [5]. It can be posited that this would potentially empower humans to retain decision-making authority. Furthermore, advances in computational fact-checking can provide tools to verify the truthfulness of LLM output. Journalists already use fact-checking to compare claims by public figures or social networks with reference corpus to verify the authenticity of information [1]. Computational fact-checking could be a valuable asset to assess *a priori* truthfulness of the answers provided by LLMs.

To understand the impact of overly confident answers provided by LLMs, as well as other manifestations of human behaviour that might arise in future LLMs, we need to understand better how people perceive the moral dimensions of LLMs and how this influences their perceptions of correctness and truthfulness. We call on the research community to further investigate the moral manifestations of LLMs and subsequently apply these insights to the design of future systems.

ACKNOWLEDGMENTS

This work is supported by the Carlsberg Foundation project ‘Algorithmic Explainability for Everyday Citizens’.

REFERENCES

[1] Oana Balalau, Simon Ebel, Théo Galizzi, Ioana Manolescu, Quentin Massonnat, Antoine Deiana, Emilie Gautreau, Antoine Krempf, Thomas Pontillon, Gérald Roux, and Joanna Yakin. 2022. Statistical Claim Checking: StatCheck in Action. In *Proceedings of the 31st ACM International Conference*

on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, Mohammad Al Hasan and Li Xiong (Eds.). ACM, 4798–4802. https://doi.org/10.1145/3511808.3557198

[2] Brian Bushard. 2023. Fake Scientific Abstracts Written By ChatGPT Fooled Scientists, Study Finds. *Forbes* (2023). https://www.forbes.com/sites/brianbushard/2023/01/10/fake-scientific-abstracts-written-by-chatgpt-fooled-scientists-study-finds/?sh=2efbccc018b6

[3] Upol Ehsan. 2023. Tweet. *Twitter* (2023). https://twitter.com/UpolEhsan/status/1618832235360313346

[4] Susan T. Fiske, Amy J.C. Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences* 11, 2 (2007), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

[5] Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sona Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *CoRR* abs/2209.14375 (2022). https://doi.org/10.48550/arXiv.2209.14375

[6] Joshua D. Greene. 2015. The rise of moral cognition. *Cognition* 135 (2015), 39–42. https://doi.org/10.1016/j.cognition.2014.11.018

[7] Paul Hayes, Ibo van de Poel, and Marc Steen. 2022. Moral transparency of and concerning algorithmic tools. *AI and Ethics* (20 Jun 2022). https://doi.org/10.1007/s43681-022-00190-4

[8] Ketan Joshi. 2023. Tweet. *Twitter* (2023). https://twitter.com/KetanJ0/status/1615646357103992834

[9] Elisa Konya-Baumbach, Miriam Biller, and Sergej von Janda. 2023. Someone out there? A study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior* 139 (2023), 107513. https://doi.org/10.1016/j.chb.2022.107513

[10] Effie Lai-Chong Law, Asbjørn Følstad, and Nena van As. 2022. Effects of Humanlikeness and Conversational Breakdown on Trust in Chatbots for Customer Service. In *Nordic Human-Computer Interaction Conference* (Aarhus, Denmark) *(NordiCHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 56, 13 pages. https://doi.org/10.1145/3546155.3546665

[11] Joonas Moilanen, Niels van Berkel, Aku Visuri, Ujwal Gadiraju, Willem van der Maden, and Simo Hosio. 2023. Supporting Mental Health Self-Care Discovery Through a Chatbot. *Frontiers in Digital Health* (2023). https://doi.org/10.3389/fdgth.2023.1034724

[12] Chris Stokel-Walker and Richard Van Noorden. 2023. What ChatGPT and generative AI mean for science. *Nature* (2023). https://www.nature.com/articles/d41586-023-00340-6

[13] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (oct 2021), 39 pages. https://doi.org/10.1145/3476068