Safety and Accountability for Large Language Model Use in HRI

Christine T. Chang*
christine.chang@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Bradley Hayes
bradley.hayes@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

ABSTRACT

Large language models (LLMs) are rapidly being incorporated into human-robot interaction (HRI) systems. Many researchers have already acknowledged some of the potential dangers of such work. However, a comprehensive method for analyzing the safety and risk implications for LLMs in HRI does not exist. This extended abstract is a call to action. First, the LLM-HRI community must commit to a safety and risk analysis methodology. Second, the community has a responsibility to advocate for safe use of LLM-HRI technology. In this dynamic world of emerging technology, process and commitment to safety will be paramount.

KEYWORDS

risk, advocacy, public policy, model cards, system cards, safety

1 INTRODUCTION

With increasing frequency large language models (LLMs) are being released, and with an uncanny sense of urgency researchers are developing interfaces to facilitate their use with robots. Google, Apple, and others [4, 11–13, 15] have introduced research demonstrating the use of LLMs with robots, frequently showcasing impressive feats that have so far eluded other approaches. Often these products are debuted to the public without an acknowledgement of the impact that they can have. While we should laud transparency, any publication of work should also acknowledge the associated risks.

An open letter was written and signed by many esteemed HRI contributors that calls for robots to not be developed or used for questionable law enforcement activities [14]. A "consensus paper" from prominent researchers in AI and robotics was shared in 2023, highlighting large-scale risks of AI in light of the recent explosion of LLMs [3]. A cautionary "Statement on AI Risk" continues to collect signatories [6]. Now we need to turn these words into actions, bound by a commitment to acknowledge the safety implications of our work. What follows in this paper is a call to action for safety and accountability when using LLMs within HRI systems.

When publishing such research, our venues should require authors to make statements regarding what the assumptions and impacts of their work may be, both positive and negative. While many deficiencies of existing LLMs are already apparent and well documented, these measures will help to proactively mitigate harms from others that have yet to be revealed. We must not only anticipate these known unknowns, but also be aware that there exist "unknown unknowns". With this foresight we can prepare, provide safeguards, and anticipate what might come next.

Thus, our call in this short paper is two-fold:

(1) A meaningful safety and risk analysis should be included with paper submissions and open research publications.
(2) HRI researchers have a social responsibility to advocate for safe use of technology that uses LLMs.

For (1), we can look to existing research on model cards to formulate an appropriate method for doing so within HRI. We propose one such method here for the sake of discussion. For (2), we not only must take responsibility for the technology that we develop and refine, but we must also take steps to inform policymakers and make the general public aware of the current state-of-the-art and its limitations. More details follow below.

2 SAFETY AND RISK ANALYSIS FOR HUMAN-LLM INTERACTION

Risk analyses feature prominently in a variety of contexts, including financial, IT, environmental, and health. Reasons for this are largely business-driven, but also have implications for local communities and society at large. However, risk analyses are less common for academic research, which rarely discusses potential societal impacts of the work beyond the focused tasks or use cases they target. As barriers between research and application shrink, this must change.

2.1 The Case for Safety and Risk Analysis

First, we must acknowledge that disparate risk tolerances and risk attitudes exist, both among researchers and among the public. The designers and developers of a system should have detailed insights into how the system functions, and thus perhaps have different perceptions of risk and safety for the performance of the technology than a user who was not involved in the system design and development. Such a user may be more risk averse, but has less insight into the functionality of the system to justify their perceptions. Formative research in public perception of safety and risk showed that new technologies invoked a higher risk rating [5]. Recent work even argues that we should make the AI system itself more adaptable to different risk attitudes [10].

Second, the external perception of emerging technologies is influenced by media portrayals and framing, perpetrating stories and images related to a "robot apocalypse," "Skynet," and other scenarios where robots (physical and computer-based) overtake humanity. It is our responsibility to interrupt this narrative when it begins to permeate the public and instead replace it with one of informed caution when necessary.

Finally, when it comes to conducting research with LLMs on robots and chatbots as they interact directly with humans, there are real and impactful implications of this work. The risks involved

*This author is a Draper Scholar. Any opinions or recommendations expressed in this work are those of the author and do not necessarily reflect the views of Draper.

HRI24, Human-LLM Interaction Workshop, Boulder, CO

ACM ISBN 978-1-4503-XXXX-X/24/03
https://doi.org/10.1145/nnnnnnn.nnnnnnn

in the results of these technologies depend on the purposes for which an LLM was designed, what safety measures have been implemented outside of the LLM, the embodiment (or not) of the robot, how the robot itself is characterized, and other factors. Implications can include something as benign as moving the wrong object on a table or something more dangerous such as a person following harmful advice as convinced by the output of one of these models.

Thus, it is imperative that as ethical researchers we should be forthcoming with the risks that the results of our research pose. One response is the "model card" or "system card," initially proposed by Mitchell et al. [7]. A model or system card details the implementation of a machine learning model or artificial intelligence system, and may also include other details like information about training data and target use cases or applications. However, the subsections relevant to risk are distributed throughout the document. Furthermore, this practice is not inclusive to cyberphysical or HRI systems. We propose to alter the model card baseline to be more relevant to robotics, and to be more thorough than an impact statement [1].

2.2 A Model (Card) for Safety and Risk Analysis

We start an outline for a safety and risk analysis by examining the model cards presented in [7] as well as the main sections of the system cards published by OpenAI for GPT-4 and GPT-3.5 [8, 9]. The system card for GPT-4 especially focuses on safety throughout the document, examining safety from the viewpoint of the user, and links to additional reports on disinformation, misuse, and impacts on the economy and labor market. These documents present workable models for what a meaningful risk analysis could look like for HRI that incorporates LLMs, in contrast with a *meaningless analysis* that does not provide functional or actionable insights.

A proposed outline for risk/safety analysis:

- **Overview** of the system and its intended purpose
- **Physical** risks (individual and large scale)
- **Psychological** risks (individual and large scale)
- **Societal** risks
- Other risks
- **Ethical** implications

These categories focus on the effects of the LLM-HRI system and could accompany a model/system card that focuses inwardly on the development of the model, its evaluation, or its intended uses, for example. Factors to consider in each category include the data used for training, testing, and validation; the impacts and risks beyond intended uses; design decisions for implementing the model with the robot; testing methods; and the composition of the development team. An acceptable level of analysis would mean that each category has received due diligence; these categories and sub-categories could be revised and detailed once put into practice.

3 ADVOCACY AS SOCIAL RESPONSIBILITY

As with any new technology or practice, at first it takes concerted action to make it a convention. Here we call on the HRI community to be part of instituting this change. To do so we must look to 3 different and parallel approaches. (1) First, we need to appeal to our colleagues in this immediate field and related ones to build a network of collaborative and influential action. (2) Second, we need to increase the visibility of our published work on LLM-HRI as well as its potential impacts to the public and the mainstream media. (3) And third, we must connect with policymakers who have the power to enact laws around the safe use of emerging technologies.

Collegial action. Within the field of HRI, and especially where LLMs are incorporated into HRI research, we have a responsibility to discuss with our colleagues the real impacts of our work and normalize openness about potential unintended effects. Effective scientific culture requires the flexibility to intellectually explore the far reaching ramifications of our results, both positive and negative. Beyond discussion, we as a community must set specific and actionable expectations for how to responsibly characterize the risks and benefits associated with publishing our work.

Media exposure. The visibility of our work and its implications is an important element of safety, risk, and accountability. Academic venues such as conferences and journals are essential, and we include those in the previous point on discourse with colleagues. However, without reaching the general public, or even the industries for which our work is intended, our efforts are meaningless. Thus, we have an obligation to reach out to media who can help amplify our work, especially regarding the safe and proper use of LLMs in HRI and HCI, with a balanced and responsible perspective. Outlets such as The Conversation [2] are a starting point, and even the media engines at our individual institutions. Local newspapers, radio stations, and podcasts provide farther-reaching and potentially impactful modalities for helping to educate the general public about the issues at hand, and for encouraging ongoing discussions and subsequent action. We must each take initiative in starting and driving the discussion, otherwise someone else will.

Policy. Our ultimate social responsibility is to convince policymakers at all levels to enact rules and laws that promote safe and responsible use of robotic and AI-based systems. In collaboration with our colleagues, and with the visibility provided by media outreach, along with the social clout of our academic work, we are uniquely positioned to work with policymakers. The most visible example of this might be tech CEOs testifying in front of Congressional committees, but we can work at the local and state levels as well. Because of the rapidly changing status of the technologies that we research, our contributions are urgently and constantly needed. Policymakers at every level and in a variety of government agencies are looking for guidance. Practically, this will mean (1) developing accessible training for researchers to effectively inform and propose realistic policy recommendations, (2) creating a community of HRI researchers to collaboratively advocate on issues, and (3) including policy recommendations as part of our safety and risk analyses. One policy change we can implement immediately is to update journal and conference submission guides to require a safety and risk analysis as part of the submission process.

4 NEXT STEPS

This extended abstract proposed a number of new and urgent actions for the Human-LLM/HRI research community to take. Our very next steps should be to update paper submission guides to require a safety and risk analysis for all relevant submissions and to initiate an advocacy group within the HRI community that can lead the way for further necessary change.

REFERENCES

[1] [n. d.]. *ICML 2024.* https://icml.cc/Conferences/2024/CallForPapers

[2] 2024. The Conversation: In-depth analysis, research, news and ideas from leading academics and researchers. https://theconversation.com/us

[3] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. 2023. Managing AI Risks in an Era of Rapid Progress. http://arxiv.org/abs/2310.17688 arXiv:2310.17688 [cs].

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. (July 2023). https://robotics-transformer2.github.io/

[5] Baruch Fischhoff, Paul Slovic, Sarah Lichtenstein, Stephen Read, and Barbara Combs. 1978. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences* 9, 2 (April 1978), 127–152. https://doi.org/10.1007/BF00143739

[6] Geoffrey Hinton. 2023. Statement on AI Risk | CAIS. https://www.safe.ai/statement-on-ai-risk

[7] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[8] OpenAI. 2020. GPT-3 Model Card. https://github.com/openai/gpt-3/blob/master/model-card.md

[9] OpenAI. 2023. GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[10] Elisabeth Paté-Cornell. 2024. Preferences in AI algorithms: The need for relevant risk attitudes in automated decisions under uncertainties. *Risk Analysis* 1, 7 (Jan. 2024). https://doi.org/10.1111/risa.14268 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.14268.

[11] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning. https://openreview.net/forum?id=wMpOMO0Ss7a

[12] SYNCED. 2023. Apple Repurposes Large Language Models for Reinforcement Learning challenges in Embodied AI | Synced. https://syncedreview.com/2023/11/01/apple-repurposes-large-language-models-for-reinforcement-learning-challenges-in-embodied-ai/

[13] Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Walter Talbott, Katherine Metcalf, Natalie Mackraz, Devon Hjelm, and Alexander Toshev. 2023. Large Language Models as Generalizable Policies for Embodied Tasks. http://arxiv.org/abs/2310.17722 arXiv:2310.17722 [cs].

[14] Tom Williams. 2020. No Justice, No Robots: An Open Letter. https://docs.google.com/forms/d/e/1FAIpQLSfDw-V67BfMGMJLZ3XeAGJonzc1bT7cJHI4JTRIHueoGMIa9Q/viewform?embedded=true&usp=embed_facebook

[15] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. 2023. Language to Rewards for Robotic Skill Synthesis. https://doi.org/10.48550/arXiv.2306.08647 arXiv:2306.08647 [cs].

Received 2 February 2024