I robot, you Jane? Ethics in the age of social robots

Erik Lagerstedt
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Gothenburg, Sweden
erik.lagerstedt@gu.se

Christine Howes
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Gothenburg, Sweden
christine.howes@gu.se

I. Introduction

Ethics is a core aspect of research, however, it is not always the core interest of the researchers. Presumably, researchers want their work and contributions to be ethical, but developing a deeper understanding of ethics might take time and effort that the researchers would prefer to spend on their core subject. To simplify the process, it can therefore be efficient to rely on experts in ethics to review research proposals or to develop guidelines (for example [1]) to adhere to. These resources can thus constitute a kind of safety measure to prevent unethical practice, while reducing the demands on ethical competence of the individual researchers. Designing systems with gates that require passing some ethical evaluation could arguably contribute toward a systematic approach to ethical research.

II. Problems With the Gates

There are, however, drawbacks to this approach of standardised checkpoints. Experts in ethics might not always have sufficient access to the particular conditions of each instance that they evaluate, and guidelines may be too general, or rely on assumptions that are obvious in some domains but may not hold in others. In such cases, the tools for ensuring ethical practices might instead prevent ethical research. The field of biomedical research has a long history of engaging with ethics, partly due to several noticeable examples in modern history of unethical large scale studies [2]. For that reason, there are now plenty of resources developed to support ethical practices of such research. When other fields have realised the need to increase the engagement with ethical perspectives, they have often been inspired by the field of medicine, however, the transferred material from the biomedical domain has sometimes only been superficially adapted to the new domain [2].

Another (yet related) problem is when efforts of generalising processes or rules lead to general rules being applied in inappropriate ways, or ways that prevent ethical practices. For example, following the conventional rules (instated with the intentions of protecting the subjects of the research study) related to research ethics when investigating vulnerable communities in on-line environments might actually be harmful for the community, and, in turn, its members (see e.g. [3]). The problem highlighted in this example is how the requirement of informed consent threatens to disrupt the sense of security felt by the community members, a sense that is critical for this kind of safe space. These rules only apply to researchers, and not other stakeholders, meaning that vulnerable or non-breaking individuals and communities might effectively be silenced or made invisible in the scientific literature. In turn, this under-documentation might make evidence based methods increasingly bad at protecting, supporting, or including the most vulnerable people. This is particularly problematic in combination with the fact that stakeholders with a potential interest in exploiting vulnerable people will be less restricted by ethical requirements.

III. Robots and Languages

To put this submission into more specific context, we will consider natural language interaction between humans and robots. A common argument for using social robots is that their human-like features will allow humans to interact with these machines in an intuitive way by relying on anthropomorphism (see e.g. [4]). This does, however, directly raise questions regarding what is considered human-like and intuitive, and indirectly raise questions regarding who is allowed to decide (and for whom) how to answer the direct questions. The power structures related to who makes the calls in the design and deployment of social robots can be quite complex [5], and the representation and visibility of different kinds of people studies of human-robot interaction is somewhat skewed (see e.g. [6]), which means that the basis for the decisions is warped. Part of the problem might come from a formulaic or overly standardised way of investigating humans, which over time enforces certain ideas of what a human can be.

This is not a new issue; it is analogous to the now widely accepted notion that insights about human psychology were based on subject populations who were "WEIRD" (Western, educated, industrialized, rich, and democratic), and not as universal as had been assumed [7]. While psychologists and linguists, for example, have been aware of this issue for some time, it is only recently that the same questions are being asked in the field of human-computer interaction [8]. The current paradigms of Large Language Models (LLMs, which are increasingly being used in social robotics studies [9]) only compound the problems. Of over 7000 languages in the world,[1] ChatGPT and other LLMs are trained on only those with a significant internet presence—which could be as little as

[1]https://www.ethnologue.com/insights/how-many-languages

2% of the world's languages,[2] and is overwhelmingly skewed to English—the paradigm case of the WEIRD.

The problems related to LLMs are carried over to the artefacts in which they are implemented. Given that part of the intuitive interaction that the human-like interface is intended to afford is natural language, the domain of social robotics might be particularly affected by this risk. In addition to bias toward English, there is a larger problem related to the the nature of natural language itself. An overly normative view of language might lead to dynamic and pragmatic aspects of language being ignored, reducing the utility of the robot. Importantly, not all humans might be affected equally by such faults in the technology [10]. For under-privileged people who often experience exclusion, the reduced utility might not appear in standardised usability evaluations, due to lower expectations among this group of people [11], [12]. It is also not only a question of reduced usability as the assumptions built into the technology might even be harmful for people breaking such norms. For instance, an overly narrow view of what English sounds like can be harmful in terms of reduced learning outcomes for children with a "non-typical" English dialect when relying on education robots [13].

IV. Some Issues to Consider

There are many aspects to consider in relation to developing the conditions for ethical research, not least given the many ways ethics can apply to, and intersect with, research. Below are several such aspects of particularly important to highlight when scrutinising the conditions for ethical research.

A. Engage With the Problem

Although there are many benefits of guidelines and checklists, such as constituting benchmarks and resource for inspiration, they will never replace the need for meaningfully engaging with the problem at hand. The study of a vulnerable on-line community mentioned above [3], was only possible investigating and engaging with the ethical concerns directly, rather than superficially fulfilling ethical requirements based on guidelines and checklists.

B. Be Careful With Statistics

Statistics are often useful tools for summarising data, providing an overview of otherwise messy or unwieldy datasets. Individual numbers can thus be used to highlight things like the locations, spreads, or sizes of clusters. These numbers are, however, properties of the dataset, not of the individual datum of which it is composed. The clarity and simplicity of the statistics are gained by considering the collective at the cost of dismissing the nuances of the individuals. To use these statistics by ascribing them back to the individuals within the dataset is therefore fundamentally problematic. There is arguably an ethical necessity of properly investigating statistical outliers to avoid perpetuating systematic exclusions [10].

[2]https://seo.ai/blog/how-many-languages-does-chatgpt-support

C. Acknowledge the Complexity

It is useful to remember that "all models are wrong but some are useful" [14, p. 2]. Similar to the arguments regarding statistics, models are simplifications made to highlight some phenomena at the cost of hiding others. To get a more complete and inclusive view of a complex system, it is necessary to assume a pluralistic stance [15]. In the context of designing relevant social robots, it is necessary to embrace diversity [16].

D. Reflect on the Power-Wielding Aspect

To wield power is to increase or reduce the range of possibilities someone else has. This can be in terms of ways to participate, actions to do, decisions to make, and more. When making design decisions regarding technological artefacts the designer determines how, and by whom, the artefact can be used [5]. Similar effects are consequences of decisions regarding the extent of deployment of the artefacts. The priorities of the different stakeholders do not always align [13], [17], making it necessary as a designer or researcher to make responsible decisions.

E. Remember the Voiceless

There are many reasons why a stakeholder might not be able to voice their concerns. It might be a human with conflicting interests, at an age or state where their opinions can be difficult to interpret, they might harbour unreasonably low expectations due to a history of oppression. The stakeholder might also not be a human. For example, humans are just one of many species of animals that might be affected by decisions regarding technology use. Given that humans are part of a larger ecosystem, it might even be relevant to consider parts of the environment stakeholders (see e.g. [18]). It is therefore particularly important for designers and researchers to pay attention to the needs of those that cannot call attention to them.

V. To Be Continued

Instead of a conclusion, this submission ends with a request to not attempt to find a final solution to this issue, but instead keep the issue open to maintain the struggle for improvement. The *process* of engaging with the specific and situated ethical nature of each instance should be the core of the solution, rather than some general artefact. Prior experience of such engagements, whether embodied in the various stakeholders, documented in guidelines, or accessible in some other way, can be important support for providing support to get further with assessment of the new situation, but they should be considered resources rather than rules.

That said, it might still be important to keep some hard, blunt rules to protect from certain exploitation. The aim should, however, not only be to refine such rules, but also to ultimately abolish them. For the latter to be a possibility, it is necessary to encourage everyone to engage with ethical issues in a meaningful way, which, in turn, provides training, experience, and improved competence in terms of assessing ethical issues.

REFERENCES

[1] World Medical Association, "Declaration of helsinki," 2018. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

[2] A. Markham, "Ethics," in *Handbook on Digital Criminology* (H. Mork Lomell and M. Kaufman, eds.), De Gruyter Press, in press.

[3] Y. H. af Segerstad, C. Kullenberg, D. Kasperowski, and C. Howes, "Studying closed communities on-line: Digital methods and ethical considerations beyond informed consent and anonymity," *Zimmer M. & Kinder-Kurlanda K.(szerk.), Internet Research Ethics for the Social Age. New Challenges, Cases, and Contexts. Bern, Switzerland: Peter Lang US*, 2016.

[4] M. F. Damholdt, O. S. Quick, J. Seibt, C. Vestergaard, and M. Hansen, "A scoping review of hri research on 'anthropomorphism': contributions to the method debate in hri," *International Journal of Social Robotics*, vol. 15, no. 7, pp. 1203–1226, 2023.

[5] K. Winkle, D. McMillan, M. Arnelid, K. Harrison, M. Balaam, E. Johnson, and I. Leite, "Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 72–82, 2023.

[6] K. Winkle, E. Lagerstedt, I. Torre, and A. Offenwanger, "15 years of (who) man robot interaction: Reviewing the H in Human-Robot Interaction," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–28, 2023.

[7] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?," *Behavioral and brain sciences*, vol. 33, no. 2-3, pp. 61–83, 2010.

[8] S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke, "How weird is chi?," in *Proceedings of the 2021 chi conference on human factors in computing systems*, pp. 1–14, 2021.

[9] T. Williams, C. Matuszek, R. Mead, and N. Depalma, "Scarecrows in oz: the use of large language models in hri," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–11, 2024.

[10] J. Rosén and E. Lagerstedt, "Speaking properly with robots," in *HRI'23 Workshop—Inclusive HRI II, Equity and Diversity in Design, Application, Methods, and Community, Stockholm, Sweden, March 13, 2023, Co-located with the 2023 International Conference on Human-Robot Interaction (HRI 2023)*, pp. 1–3, 2023.

[11] A. Pyae and P. Scifleet, "Investigating differences between native english and non-native english speakers in interacting with a voice user interface: A case of google home," in *Proceedings of the 30th Australian conference on computer-human interaction*, pp. 548–553, 2018.

[12] J. Kristensen and J. Lindblom, "How young people living with disability experience the use of assistive technology," in *International Conference on Human-Computer Interaction*, pp. 250–268, Springer, 2021.

[13] D. K. Singh, M. Kumar, E. Fosch-Villaronga, D. Singh, and J. Shukla, "Ethical considerations from child-robot interactions in under-resourced communities," *International Journal of Social Robotics*, vol. 15, no. 12, pp. 2055–2071, 2023.

[14] G. E. Box, "Robustness in the strategy of scientific model building," in *Robustness in statistics*, pp. 201–236, Elsevier, 1979.

[15] S. D. Mitchell, *Unsimple truths: Science, complexity, and policy*. University of Chicago Press, 2009.

[16] D. Vernon, "An african perspective on culturally competent social robotics: Why dei matters in hri," *IEEE Robotics and Automation Magazine*, in press.

[17] H. L. Bradwell, R. Winnington, S. Thill, and R. B. Jones, "Ethical perceptions towards real-world use of companion robots with older people and people with dementia: survey opinions among younger adults," *BMC geriatrics*, vol. 20, pp. 1–10, 2020.

[18] M. P. de La Bellacasa, *Matters of care: Speculative ethics in more than human worlds*, vol. 41. U of Minnesota Press, 2017.