frontiers

Editorial: Responsible Robotics

Martim Brandão [1,*], **Masoumeh Mansouri** [2] **and Martin Magnusson** [3]

[1] *Department of Informatics, King's College London, London, United Kingdom*
[2] *School of Computer Science, University of Birmingham, Birmingham, United Kingdom*
[3] *School of Science and Technology (AASS), Örebro University, Örebro, Sweden*

Correspondence*:
Martim Brandão
martim.brandao@kcl.ac.uk

Keywords: Robotics, Responsible Innovation, Responsible Robotics, Trustworthy Robotics, Critical Robotics, AI and Society, Robot Ethics

RESPONSIBLE AI AND ROBOTICS

Recent work in both academia, industry, and journalism has brought widespread attention to various kinds of harmful impact that AI can have on society. These are very often concentrated on marginalized social groups. AI algorithms may unintentionally reinforce social prejudice Bolukbasi et al. (2016) and biased conceptions of gender Adams and Loideáin (2019); Hamidi et al. (2018), race Sweeney (2013), age Rosales and Fernández-Ardèvol (2019) or disabilities Guo et al. (2020), they may lead to unfair access to opportunities Dastin (2018); Angwin et al. (2016), discriminatory pricing practices Bar-Gill (2019); Hannak et al. (2014), etc. Recent work has also shown that many seemingly technical issues in machine learning are actually socio-technical. For example: the over-fitting of machine learning models, the choice of dataset or learning objective, and other aspects of learning may lead to algorithms performing poorly on unrepresented or unmodeled groups of people Brandao (2019); Barocas et al. (2019); Buolamwini and Gebru (2018). A growing community of Fairness, Accountability, Transparency, and Ethics of AI[1] is now approaching these topics from a socio-technical point-of-view, in order to identify, understand, and alleviate such issues.

Robotics, as a technology focused on automation and intelligent behavior, also abounds in similar ethical and social issues that need to be identified, characterized, and considered in design. While many of the same problems with AI will also be present in robotics, the physical nature of robotics raises new aspects of the social and ethical nature of these technologies. As one example: models that are considerably less accurate on certain groups of people can lead to physical safety differentials Brandao (2019), where robots or autonomous vehicles using those models are more likely to collide with those groups. Additionally, there are physical safety concerns with respect to surgical and other medical robots Yang et al. (2017); Ficuciello et al. (2019), as well as concerns of physical and political security—not least concerning autonomous weapon systems and the dual-use of robot technologies like autonomous cars and drones Brundage et al. (2018); Sparrow (2007).

The physical design and visual appearance of robots also introduce new aspects to responsible development. For example, people's moral evaluation of robot decisions can be affected by whether

[1] Example venues: ACM Conference on Fairness, Accountability, and Transparency (FAccT), AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)

the robot is more or less human-like Malle et al. (2016), the design of robots in a care setting affects caregivers and caretakers van Wynsberghe (2021); Kubota et al. (2021), the choice of sensors, measurements and motion has an impact of privacy Calo (2011); Eick and Antón (2020); Luo et al. (2020), and the ethics of deception takes on new shape Danaher (2020).

The robotics community has been discussing ethics for long[2]. Recent workshops have also started bringing attention to philosophical problems in robotics[3] and issues such as bias[4] and transparency[5]. These efforts share a common goal of developing robotics technologies responsibly—they are part of "Responsible Robotics" or "Trustworthy Robotics".

A similar effort on "Critical Robotics" Serholt et al. (2021) has focused on questioning current practices in robotics research. These range from how older adults are represented in HRI Burema (2021) and ethical issues in education robots Serholt et al. (2017), to normative dimensions of speech used by researchers Brandao (2021), their technological optimism Šabanović (2010) and the influence of their social background in research directions Forsythe (2001); Šabanović (2010).

THIS RESEARCH TOPIC

This research topic gathers a diverse set of articles on Responsible Robotics. They range from user studies and philosophical inquiry, to modeling, algorithmic, and governance methods. Our goal when organizing this research topic was exactly to join various approaches in a single edition—to allow for greater multidisciplinary exchange under the common mission of Responsible Robotics. We believe that Responsible Robotics should focus both on *identifying* social and ethical issues, and on *designing* methods to account for (and alleviate) such issues—thus the focus of this edition on both understanding and *acting* on social and ethical issues.

Two articles in the research topic are focused on eliciting social and ethical issues *from users and stakeholders*. Lutz and Tamò-Larrieux (2021) investigate privacy concerns of lay users and their impact on technology use intentions, when using social robots that are either privacy-friendly or privacy-invasive (e.g. listen to conversations, share data with third parties). Colombino et al. (2021) use ethnographic studies, interviews and futuristic autobiographies to identify organizational principles, potential roles, and ethical design considerations for a robot that collaborates with disabled employees.

Three articles are more focused on methods, or socio-technical solutions to ethical problems in robotics. Webb et al. (2021), for example, focus on methods for conducting investigations of accidents involving humans and robots. In particular, they propose and preliminarily evaluate a role-play-based methodology for investigating accidents, and to evaluate the testimonies that humans can give in forensic investigations of such accidents. Hurtado et al. (2021) focus on issues of harmful social bias in robot learning and how they could be detected and alleviated. Namely, they show through various examples how social robot navigation techniques that mimic human behavior may lead to harmful behavior, such as higher intrusion of personal space or longer waiting times for some groups compared to others. Winfield et al. (2021) focus on issues of transparency from a governance perspective. They describe a new draft standard on transparency for autonomous systems, with several contributions such as transparency levels, measurability, stakeholders, and example-based guidance on using the draft standard.

[2] ICRA 2007/2009/2011 workshops on Roboethics, ICRA 2014 workshop on "Robotics and Military Applications"

[3] Robophilosophy Conference

[4] ICRA 2019 workshops on "Bias-sensitizing robot behaviours" and "Unlearning biases in robot design"

[5] HRI 2022 workshop on "Fairness and Transparency in HRI", ICRA 2020 workshop "Against robot dystopias"

We then dive into philosophical inquiry and frameworks for robot ethics. Rhim et al. (2021) combine work in moral philosophy and psychology to propose a model that explains human decision-making in moral dilemmas involving autonomous vehicles. Pirni et al. (2021) consider aspects of autonomy and vulnerability in the ethics of designing care robots. And Kuipers (2022) argues that AI and robotics technologies rely heavily on over-simplified models, and that the widespread use of such models can lead to the erosion of trust and cooperation effectiveness. The article can serve as an argument for why more attention should be given to the *modeling* of complex socio-technical factors in AI/robotics.

Finally, two articles in the research topic dive into issues of jobs and economics in robotics and automation. Studley (2021) argues that we should consider how robotics impacts global supply chains, international development, and global economic disparities. Kyvik Nordås and Klügl (2021) then use modeling to understand the uptake of automation technologies and its relationship with unemployment and engineering, consultancy, and manufacturing jobs. The authors use this analysis to suggest an automation policy focus on user costs and education.

We believe that the contributions collected in this special issue can be relevant to roboticists, AI practitioners, policy makers and any other stakeholders concerned with the societal impacts of AI and robotics. We hope this special issue will stimulate future work on responsible robotics.

We end with an important remark. While the abundance of social and ethical issues raised in this editorial and this Research Topic might feel overwhelming or hopeless, we believe the opposite is the case. Responsible Robotics is about clearly identifying potential issues, because by doing so it is also possible to work towards responsible methods that mitigate them. This ultimately facilitates the application of robotics and AI in ways that increase safety, efficiency, and wellbeing in many areas of life: transportation, healthcare, work life, just to name a few.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Adams, R. and Loideáin, N. N. (2019). Addressing indirect discrimination and gender stereotypes in ai virtual personal assistants: the role of international human rights law. *Cambridge International Law Journal* 8, 241–257

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks.

Bar-Gill, O. (2019). Algorithmic price discrimination when demand is a function of both preferences and (mis) perceptions. *University of Chicago Law Review* 86

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning* (fairmlbook.org). `http://www.fairmlbook.org`

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29

Brandao, M. (2019). Age and gender bias in pedestrian detection algorithms. In *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision, CVPR*

Brandao, M. (2021). Normative roboticists: the visions and values of technical robotics papers. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 671–677. doi:10.1109/RO-MAN50785.2021.9515504

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, eds. S. A. Friedler and C. Wilson (New York, NY, USA: PMLR), vol. 81 of *Proceedings of Machine Learning Research*, 77–91

Burema, D. (2021). A critical analysis of the representations of older adults in the field of human–robot interaction. *AI & Society* , 1–11

Calo, R. (2011). Robots and privacy. *Robot Ethics: The Ethical and Social Implications of Robotics*

Colombino, T., Gallo, D., Shreepriya, S., Im, Y., and Cha, S. (2021). Ethical design of a robot platform for disabled employees: Some practical methodological considerations. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.643160

Danaher, J. (2020). Robot betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology* 22, 117–128

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics* (Auerbach Publications). 296–299

Eick, S. and Antón, A. I. (2020). Enhancing privacy in robotics via judicious sensor selection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 7156–7165. doi:10.1109/ICRA40945.2020.9196983

Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., and Siciliano, B. (2019). Autonomy in surgical robots and its meaningful human control. *Paladyn, Journal of Behavioral Robotics* 10, 30–43. doi:doi:10.1515/pjbr-2019-0002

Forsythe, D. (2001). *Studying those who study us: An anthropologist in the world of artificial intelligence* (Stanford University Press)

Guo, A., Kamar, E., Vaughan, J. W., Wallach, H., and Morris, M. R. (2020). Toward fairness in ai for people with disabilities: a research roadmap. *ACM SIGACCESS Accessibility and Computing* , 1–1

Hamidi, F., Scheuerman, M. K., and Branham, S. M. (2018). Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13

Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. (2014). Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference*. 305–318

Hurtado, J. V., Londoño, L., and Valada, A. (2021). From learning to relearning: A framework for diminishing bias in social robot navigation. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.650325

Kubota, A., Pourebadi, M., Banh, S., Kim, S., and Riek, L. (2021). Somebody that i used to know: The risks of personalizing robots for dementia care. *Proceedings of We Robot*

Kuipers, B. (2022). Trust and cooperation. *Frontiers in Robotics and AI* 9. doi:10.3389/frobt.2022.676767

Kyvik Nordås, H. and Klügl, F. (2021). Drivers of automation and consequences for jobs in engineering services: An agent-based modelling approach. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.637125

Luo, Y., Yu, Y., Jin, Z., Li, Y., Ding, Z., Zhou, Y., et al. (2020). Privacy-aware uav flights through self-configuring motion planning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 1169–1175. doi:10.1109/ICRA40945.2020.9197564

Lutz, C. and Tamò-Larrieux, A. (2021). Do privacy concerns about social robots affect use intentions? evidence from an experimental vignette study. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.627958

Malle, B. F., Scheutz, M., Forlizzi, J., and Voiklis, J. (2016). Which robot am i thinking about? the impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE), 125–132

Pirni, A., Balistreri, M., Capasso, M., Umbrello, S., and Merenda, F. (2021). Robot care ethics between autonomy and vulnerability: Coupling principles and practices in autonomous systems for care. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.654298

Rhim, J., Lee, J.-H., Chen, M., and Lim, A. (2021). A deeper look at autonomous vehicle ethics: An integrative ethical decision-making framework to explain moral pluralism. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.632394

Rosales, A. and Fernández-Ardèvol, M. (2019). Structural ageism in big data approaches. *Nordicom Review* 40, 51–64

Šabanović, S. (2010). Robots in society, society in robots. *International Journal of Social Robotics* 2, 439–450

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., et al. (2017). The case of classroom robots: teachers' deliberations on the ethical tensions. *AI & Society* 32, 613–631

Serholt, S., Ljungblad, S., and Ní Bhroin, N. (2021). Introduction: special issue—critical robotics research. *AI & Society* , 1–7

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy* 24, 62–77

Studley, M. (2021). Onshoring through automation; perpetuating inequality? *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.634297

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM* 56, 44–54

van Wynsberghe, A. (2021). Social robots and the risks to reciprocity. *AI & Society* , 1–7

Webb, H., Dumitru, M., van Maris, A., Winkle, K., Jirotka, M., and Winfield, A. (2021). Role-play as responsible robotics: The virtual witness testimony role-play interview for investigating hazardous human-robot interactions. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.644336

Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., et al. (2021). Ieee p7001: A proposed standard on transparency. *Frontiers in Robotics and AI* 8. doi:10.3389/frobt.2021.665729

Yang, G.-Z., Cambias, J., Cleary, K., Daimler, E., Drake, J., Dupont, P. E., et al. (2017). Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy. *Science Robotics* 2, eaam8638. doi:10.1126/scirobotics.aam8638