Philosophy & Technology (2022) 35:97
https://doi.org/10.1007/s13347-022-00583-7

EDITOR LETTER

Introduction to the Topical Collection on AI and Responsibility

Niël Conradie[1,2] · Hendrik Kempt[1,2] · Peter Königs[3]

Received: 19 September 2022 / Accepted: 19 September 2022 / Published online: 2 November 2022

The rapid progress in the research and development of artificial intelligence (AI) is radically changing many aspects of our lives. Sophisticated AI is becoming common place in such diverse domains as transportation (self-driving cars, delivery drones), the workplace (work robots, automation), healthcare (medical AI, care robots), law enforcement (predictive policing, surveillance), the military (autonomous weapons systems), and entertainment (video games, sex robots). Along with these changes come moral challenges. Crucial among these challenges are those that call attention to the myriad potential relationships between AI technologies and moral responsibility. One way of making sense of the conceptual landscape is to construe it as organized around two interrelated questions about responsibility:

(1) Can we develop *Responsible AI*, and what would such an AI look like?
(2) Who, if anybody, is *Responsible for AI*: how should responsibility for the outcomes produced by AI systems be distributed?

With these two questions as organizational focal points, this topical collection seeks to advance our understanding of the ethical and social implications of AI across different fields of application.

Responsible AI has become a catchall term for the ethical design of AI systems. This has predominantly manifested in the formulation of guidelines for responsible AI. These have been promulgated by various academics, corporations, and political actors (Floridi, 2018; OECD, 2019; IEEE, 2019; Vöneky, 2020; EC, 2020), though

✉ Peter Königs
p.j.konigs@uu.nl

Niël Conradie
niel.conradie@humtec.rwth-aachen.de

Hendrik Kempt
hendrik.kempt@humtec.rwth-aachen.de

[1] Applied Ethics Group, RWTH Aachen University, Aachen, Germany

[2] Philosophy and Law Department, Frankfurt School of Finance & Management, Frankfurt, Germany

[3] Ethics Institute, Utrecht University, Utrecht, The Netherlands

there is good reason to think that meeting the challenge of designing responsible AI cannot be accomplished merely through the provision of a checklist of steps to be ticked off (Dignum, 2019; Kiran et al., 2015). Taken more widely, the concern with responsible AI can be understood as a forward-looking responsibility on those involved in the funding, development, and deployment of these systems: a duty to ensure that the AI technologies they bring forth meet certain ethical criteria. Though most (but not all) participants in the discussion can broadly agree on some elements of these criteria — e.g., non-maleficence, transparency and privacy, the provision of fair and just outcomes for stakeholders — there is substantial disagreement on how these elements should be understood and what other elements there may be (Ghallab, 2019; Jobin et al., 2019; Wang et al., 2019). Furthermore, there is often an apparent gulf between the general principles that inform the criteria and the details of how they are to be implemented technically (Hagendorf, 2020; Peters et al., 2020). Spanning this gulf requires a thorough investigation of the actual features that these technologies possess or will possess. As examples, three features frequently identified as ethically salient, though not by any means universally present in all AI systems, are some degree of autonomy in operation (Franklin & Graesser, 1997; Sparrow, 2007; Gunkel, 2019; Nyholm, 2020), internal complexity resulting in the possible opacity of its workings and thus the correct interpretation of its outputs (Castelvecchi, 2016; Holm, 2019; Pedreschi et al., 2019), and communicative or discursive abilities (Gunkel, 2018; Guzman & Lewis, 2019). Each of these features gives rise to unique ethical considerations. Viewed in this way, the challenges facing the design of responsible AI can be taken as twofold. First, the challenge of arriving at the normatively appropriate principles and deriving the subsequent criteria. Second, the challenge of how these criteria should be implemented given the actual features of contemporary and future AI technologies. Although much insightful work has been undertaken regarding both these challenges, there remains much to be done. As examples: should responsible AI account for apparent moral pluralism, and if so how? How are considerations of system efficiency and the need for transparency and explainability to be weighed against each other in cases where there is an unavoidable tradeoff? And, since stakeholder trust is vital for the successful implementation of AI, what role should trust play in the ethical design of these systems?

The second, related question concerns *Responsibility for AI*, or how we attribute responsibility for what an AI does, on its own or in collaboration with human agents. One pressing worry among AI scholars is the difficulty of identifying who is responsible for the actions (or perhaps mere behaviors) of an AI. Thanks to advanced machine learning techniques, intelligent machines are approaching a degree of sophistication and autonomy that makes it difficult to fully understand, let alone predict their behavior. For this reason, it has struck many as inappropriate to assign responsibility for an AI's actions to the human agents who have causally contributed to its actions, such as its operator or its engineers (Danaher, 2016; Matthias, 2004; Sparrow, 2007). Whether autonomous AI really gives rise to such "techno-responsibility gaps", what might be problematic about them, and how they ought to be dealt with has been the subject of intense debate (see e.g. Gunkel, 2020; Johnson, 2015; Köhler et al., 2018; Königs 2022; Nyholm, 2018). Some have opposed the use of autonomous AI on these grounds, maintaining that there is something deeply objectionable about causing harm for which nobody can justly be held to account, especially in a military setting (Roff, 2013; Sparrow, 2007). Others have taken a more optimistic view, either by denying that uniquely techno-responsibility gaps emerge in the first place or by exploring strategies of bridging the gap in one way or another (for examples of the former response, see Köhler et al., 2018; Robillard, 2018; Tigard, 2021, and for the latter: Champagne & Tonkens, 2015; Himmelreich, 2019; Kempt & Nagel, 2021; Nyholm, 2018). A futuristic solution would be to treat AIs as possible bearers of responsibility in their own right precisely in recognition of their autonomous nature (Hellström, 2013). As AI becomes increasingly complex, questions surrounding the allocation of responsibility will become more significant. While the ongoing debate has provided some initial answers to these questions, many aspects remain poorly understood, including, for instance, the moral significance of responsibility gaps in non-military contexts and the very nature of AI agency.

It should not be thought that the debate is neatly aligned with our organizational distinction between *Responsible AI* and *Responsibility for AI*. Attempts to tackle how we can distribute responsibility in cases involving AI systems have necessary implications for what would constitute the ethical design of these same systems. Likewise a commitment to a certain idea of what constitutes ethical design will impact the possible attributions of responsibility to relevantly involved human agents. The interconnectedness of these two dimensions of the problem is reflected in the contributions to this Topical Connection.

Rosalie Waelen and **Michał Wieczorek** look at the issue of responsible AI through the lens of Axel Honneth's theory of recognition. They build their argument around the idea that the various manifestations of gender bias in AI systems can be understood as constituting instances of misrecognition of women in the Honnethian sense. Realizing responsible AI, however, requires more than just technological fixes to the problem of biases in AI systems. For these biases are symptoms of underlying structural injustices, which need to be addressed if one is to achieve an adequate recognition of women.

It is to the nature of AI agency that **Elena Popa** turns her attention, presenting an argument critical of attributing moral responsibility to artificial agents. Taking key insights from action theory as her foundation, she contends that only the adoption of a teleological approach can make sense of artificial agency and its dependence on human goals and values. Building on this, once the roles played by these goals and values are properly understood — as well as the incapacity of extant artificial systems to set their own goals — the conclusion is that attributions of moral responsibility to artificial agents themselves are illegitimate. However, fears of a responsibility gap can be alleviated by adopting a more complex picture of the relations at play, illustrated by a discussion of two of Matthias' seminal cases.

Another contribution to the question of whether we can reasonably ascribe artificial agents responsibility comes from **Mihaela Constantinescu**, **Constantin Vică**, **Radu Uszkai**, and **Cristina Voinea**. In reference to Aristotle's virtue ethics, they operationalize the ascription of moral responsibility through a four-feature moral responsibility test that can determine whether an artificial agent can be morally responsible. The authors apply this test to the concept of artificial moral advisors (AMAs) and conclude that those fail the test. Thus, we ought to prevent such artificial advisors to be bearers of responsibility. Luckily, we still can use AMAs if we restrict their use to enhancing our moral knowledge rather than relying on them being responsible.

Turning more directly to responsibility gaps, the dominant view in the literature is that such gaps are a problem that needs fixing. **Kevin Baum**, **Susanne Mantel**, **Eva Schmidt**, and **Timo Speith's** paper is one of two contributions to the topical collection offering novel solutions to this problem. Focusing on decision support systems, Baum et al. argue that adequate responsibility attribution requires, first and foremost, a human in the loop. But for a human in the loop to be responsible, she must meet the epistemic condition on responsibility. Drawing on action theory, they suggest that this presupposes access to the intelligent system's motivating reasons. An alternative solution to the problem of responsibility gaps is put forth by **Johannes Himmelreich** and **Sebastian Köhler**, who champion the idea that the dispute about the distribution of responsibility for an AI system's harmful outcomes is best approached as a conceptual engineering problem. Though clear from the outset that their aim is programmatic rather than fully solutionary, Himmelreich and Köhler see in this an opportunity for reformative evaluations of our current conceptual choices in the discussion about responsibility gaps. They illustrate how this is to be undertaken by identifying a list of functions the concept of responsibility plays and assessing the conceptual choices made by some views in the literature in the light of these functions. The answer to whether there are responsibility gaps, it seems, might be "it depends".

John Danaher, contradicting the dominant view in the literature, suggests that responsibility gaps might in fact be something to be welcomed, at least sometimes. For they allow us to delegate difficult moral choices to machines and thus to reduce the psychological distress typically associated with such choices. AI-generated gaps in responsibility provide a low-cost coping mechanism for dealing with the inevitable tragedies of moral decision-making.

Mario Verdicchio and **Andrea Perin** focus on the practical issue of attributing responsibility within complex medical interactions. While this topic has gained more attention over the last years, their contribution is focused on the ethical and regulatory responsibility between doctors, engineers, and technological artifacts under the light of notoriously hard-to-determine causal effects. With three requirements for regulations of use, based on the rights of patients for informed decisions, the duty of doctors to protect their patients' health, and the possibility for doctors to rely on the technology they are using, they propose a way of distributing responsibility that avoids a growing reliance of opaque decision-making systems and places responsibility firmly in human hands.

The ever growing sophistication of AI and, hence, their integration into human practices will pronounce and sharpen questions of both responsible AI and responsibility for AI on practical levels. The articles published with this Topical Collection make diverse yet equally important contributions to interpret and handle present and future questions of AI and responsibility.

Funding Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612; ELSA-AID, Bundesministerium für Bildung und Forschung (BMBF) (German Federal Ministry of Education and Research), Funding Number 01GP1910; INVIDAS, Bundesministerium für Bildung und Forschung (BMBF) (German Federal Ministry of Education and Research), Funding Number 16SV8537.

References

Castelvecchi, D. (2016). The black box of AI. *Nature, 538*, 20–23.

Champagne, M., & Tonkens, R. (2015). Bridging the responsibility gap in automated warfare. *Philosophy & Technology, 28*(1), 125–137.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way.* Springer Nature Switzerland.

EC. (2020). *On artificial intelligence - A European approach to excellence and trust*. Available: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Accessed 25 Oct 2022.

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology, 31*, 1–8.

Franklin, S., Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, and N. R. Jennings. (Eds.), *Intelligent agents III Agent theories, architectures, and languages*. ATAL 1996. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1193. Springer.

Ghallab, M. (2019). Responsible AI: Requirements and challenges. *AI Perspectives*, *1*(3), 1–7.

Gunkel, D. J. (2019). *An introduction to communication and artificial intelligence*. Polity Press.

Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*, 307–320.

Gunkel, D. J. (2018). Ars Ex Machina: Rethinking responsibility in the age of creative machines. In A. Guzman (Ed.), *Human-machine communication. Rethinking communication, technology, and ourselves* (pp. 221–236). Peter Lang.

Guzman, A. G., & Lewis, S. C. (2019). Artificial intelligence and communication: A human–machine communication research agenda. *New Media and Society, 22*(1), 70–86.

Hagendorf, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120.

Hellström, Th. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99–107.

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*(3), 731–747.

Holm, E. A. (2019). In defence of the black box: Black box algorithms can be useful in science and engineering. *Science, 362*(6425), 26–27.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*, First Edition. IEEE. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 289–299.

Johnson, D. (2015). Technology with no human responsibility. *Journal of Business Ethics, 127*(4), 707–715.

Kempt, H., & Nagel, S. K. (2021). Responsibility, second opinions and peer-disagreement: Ethical and epistemological challenges of using AI in clinical diagnostic contexts. *Journal of Medical Ethics, 248*, 222–229.

Kiran, A. H., Oudtshoorn, N., & Verbeek, P.-P. (2015). Beyond checklists: Toward an ethical-constructive technology assessment. *Journal of Responsible Innovation, 2*(1), 6–19.

Köhler, S., Roughley, N., & Sauer, H. (2018). Technologically blurred accountability. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Diebel (Eds.), *Moral agency and the politics of responsibility* (pp. 51–68). Routledge.

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, 24(3), 1–11.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics, 24*(4), 1201–1219.

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishing Group.

OECD. (2019). *Recommendation of the council on artificial intelligence*. OECD/LEGAL/0449.

Pedreschi, D., Gianotti, F., Guidotti, R., Monreale, A., & Ruggieri, S. (2019). Meaningful explanations of black box AI decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 9780–9784.

Peters, D., Vold, K., & Calvo, R. A. (2020). Responsible AI- Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society, 1*(1), 34–48.

Robillard, M. (2018). No such thing as killer robots. *Journal of Applied Philosophy, 35*(4), 705–717.

Roff, H. M. (2013). Killing in war: Responsibility, liability, and lethal autonomous robots. In F. Allhoff, N. G. Evans, & A. Henschke (Eds.), *Routledge handbook of ethics and war: Just war theory in the twenty-first century* (pp. 352–364). Routledge.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Tigard, D. R. (2021). There Is No Techno-Responsibility Gap. *Philosophy & Technology, 34*, 589–607.

Vöneky, S. (2020). Key elements of responsible artificial intelligence - Disruptive technologies, dynamic law. *Ordnung der Wissenschaft*.

Wang, Y., Olya, H., and Xiong, M. (2019). Toward an understanding of responsible artificial intelligence practices. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Maui, Hawaii, USA.