Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P665-cd

Building Resilient Governance Frameworks for Human-Robot Collaboration: Towards a More Interdisciplinary Understanding of Risk and Ethics in European Regulation

Naira López Cañellas

School of Media, Technological University of Dublin, Ireland. E-mail: D22124679@mytudublin.ie

Prof. Dr. Aphra Kerr

Department of Sociology, National University of Ireland Maynooth, Ireland. E-mail: aphra.kerr@mu.ie

Dr. Brian Vaughan

School of Media, Technological University of Dublin, Ireland. E-mail: brian.vaughan@tudublin.ie

This paper explores a wide range of perspectives on the study of AI applications and robotics offered by both technical and social disciplines, and pools them together to best capture their associated risk and safety concerns. After a comprehensive thematic review, the paper concludes that finding the right approach to regulating them requires both an interdisciplinary point of view and the interrogation of the underlying narratives surrounding human-robot interaction. This research contributes to current EU-wide efforts to enact a legislative framework that both supports innovation and protects citizens' rights in the domain of emerging technologies.

Key words: human factors, robotics, artificial intelligence, interdisciplinary research, ethics, risk

1. Introduction

Researchers of the technical and social dimensions of technology face an impossible dilemma: when a novel technology is in its early stages of development, it is often hard to anticipate its repercussions (Collingridge, 1980 as cited in van de Poel, 2020). However, failing to do so in time might result in a lock-in effect, making it hard to tackle any unintended consequences once said technology has become embedded (idem). EU policy circles seeking to regulate Artificial Intelligence (AI) and robotics have tried to reconcile this tension by combining two complementary outlooks: a consequentialist approach – whereby risk gradients determine each product's requisites for compliance, as in the Artificial Intelligence Act (AIA) –, and a virtue-based approach – often crystallized in ethics guidelines, as found in Hagendorff (2022).

2. Methods

To capture how said regulatory trends shape both novel technologies and the society around them, this research draws upon a thematic review of 40 peer-reviewed articles on the topic of AI policy and ethics. These have been selected from a grand total of 2753 relevant entries found in the database Web of Science focusing on either consequentialist or virtue-based assessments of AI-powered devices in human-robot collaboration (HRC) in the workplace. A joint analysis of these publications reveals considerably dissenting understandings of the risks and benefits of HRC, depending on which aspects they highlight – e.g. novelty, expressiveness, exposition time or degree of anthropomorphism— and the values they uphold – e.g. efficiency, reliability, cost minimization or sustainability (Vasilescu & Filzmoser, 2021).

These elements are selected for analysis because they conform the basis of each author's interpretation of the relationship between the social and technical aspects of HRC, which in turn explains the regulatory and ethical framework they implicitly or explicitly invite (van de Poel, 2020; van Berkel, 2022). Far from anecdotal, such conceptualizations are often 'constitutive and performative', can act as self-fulfilling prophecies, and have the power of 'legitimizing, (…) slowing or speeding up certain innovations (…) by the most powerful actors' (Vicsek, 2021: 6). Thus, finding the most suitable regulatory approach to minimize the irresponsible use of data-driven applications starts with examining these underlying narratives.

3. General Trends

The research finds that most of the reviewed articles align with the value-sensitive design (VSD) endorsed by EU institutions. This approach posits the need to embed a core set of values into new technology, usually a variant of 'Weberian principles at the core of public sector bureaucracies' such as 'transparency, equality, democratic oversight, and safeguarding citizens' well-being' (Willems et al., 2022: 2).

Umbrello argues that this view represents an attempt at shifting the attention towards 'technical means to operationalize' the response to public concerns, without questioning whether the harm could be prevented, or why it happened in the first place (2020: 18). Similarly, Vampley et al. (2018) claim that relying on these values is insufficient (e.g. transparency doesn't guarantee any significant reduction of AI applications' harmful effects), and hardly implementable in practice. Instead, several authors suggest concentrating on enhancing human agency and autonomy by asking 'what social visions technologies serve' (Vicsek, 2021: 13) and seeking to embed further constraints to ensure AI design doesn't only consider 'the optimal end result, but also acceptable ways to achieve this goal' (van Berkel et al., 2022: 2). Similarly, they urge regulators to be aware of many AI applications' track

record of taking undesirable shortcuts such as 'changing users' preferences so that they are more predictable' (Whittlestone et al., 2021: 1012).

4. Ways Forward

Multiple articles issue a shared recommendation that is of relevance here: there is insufficient exchange between stakeholder groups in the field (from civil society and interest group representatives to multilevel institutions, the private sector and standard-setting organizations), and even between professionals of the same sector (Whittlestone et al., 2021; Brynjolfsson, 2022). While calls for interdisciplinary approaches to policymaking, research and technology design are widespread, detailed guidelines on how to carry this out in practice or ambitious experiments on the matter remain scarce. Van Berkel et al. (2022) argue that it is precisely this disconnect between those concerned with the 'technical' and the 'social' side of technology that is responsible for many of the biases, inefficiencies and responsibility gaps associated with human-robot collaboration, thus justifying the need for urgent action on this front.

Equally as widespread in the sources reviewed are fears that robots will soon curtail the agency of workers, consumers and even by-standers. The perspective of large-scale deployment of robots in the workplace is commonly associated with negative psychological effects on workers (Vasilescu & Filzmoser, 2021) and fears of downward pressure on working conditions, a rise in inequality, and unemployment (idem).

In response, Brynjolfsson (2022) proposes moving from *automation* to *augmentation* to avoid concentrating knowledge and power predominantly in the side of technology, and designing robots (or, more appropriately, *cobots)* with the aim of making them more complementary to humans, prioritizing each sides' capabilities and fair task allocation instead of cost-cutting and efficiency (idem). This necessitates that trade unions are involved in both the design and decision-making processes in the sector (Vicsek, 2021), so that each worker retains their agency to determine how they interact with their technological counterparts (idem).

4. Conclusion

The implications of this research are multifold. First, it becomes apparent that each actor's portrayal of the relationship between the social and technical aspects of HRC serves to contextualize their regulatory proposal. Second, we find that while the prevailing view in both the European private and public sectors support some version of VSD, critics point out that this approach proves insufficient at addressing the harms that come with the use of AI applications and robotics. Thirdly, it is concluded that while most authors believe interdisciplinary research, policymaking and design to be the most viable pathways to address the shortcomings of widespread AI and robotics deployment, most activity in the field is far from incorporating such perspective in practice. Further research is needed to anticipate the implications that these legal and ethical governance trends will have on the future of most economic sectors, as they all progressively come to rely on AI applications and robotics to carry out more of their critical operations.

References

Brynjolfsson, Erik. "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence" *Daedalus:* 2022; 151 (2): 272–287. doi: https://doi.org/10.1162/daed_a_01915

Hagendorff, Thilo. "A Virtue-Based Framework to Support Putting AI Ethics into Practice" *Philosophy of Technology,* 35, 55 (2022). https://doi.org/10.1007/s13347-022-00553-z

Umbrello, Steven. "Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design" Science and Engineering Ethics 26, 575–595 (2020). https://doi.org/10.1007/s11948-019-00104-4

Vamplew, Peter; Dazeley, Richard; Foale, Cameron; Firmin, Sally & Mummery, Jane. "Human-aligned artificial intelligence is a multiobjective problem" *Ethics Information Technology,* 20, 27–40 (2018). https://doi.org/10.1007/s10676-017-9440-6

van Berkel, Niels; Tag, Benjamin; Goncalves, Jorge and Hosio, Simo. "Human-centred artificial intelligence: a contextual morality perspective", *Behaviour & Information Technology* (2022), 41:3, 502-518, DOI: 10.1080/0144929X.2020.1818828

van de Poel, Ivo. "Three philosophical perspectives on the relation between technology and society, and how they affect the current debate about artificial intelligence" *Human Affairs,* 2020: 30(4), 499-511. https://doi.org/10.1515/humaff-2020-0042

Vasilescu, Dragos-Cristian and Filzmoser, Michael. "Machine invention systems: a (r)evolution of the invention process?" *AI & Society* 36, 829–837 (2021). https://doi.org/10.1007/s00146-020-01080-1

Vicsek, Lilla. "Artificial intelligence and the future of work – lessons from the sociology of expectations" *International Journal of Sociology and Social Policy*, 2020: Vol. 41 No. 7/8, pp. 842-861. https://doi.org/10.1108/IJSSP-05-2020-0174

Willems, Jurgen; Schmid, Moritz J.; Vanderelst, Dieter; Vogel, Dominik & Ebinger, Falk. "AI-driven public services and the privacy paradox: do citizens really care about their privacy?" *Public Management Review*, 2022. DOI: 10.1080/14719037.2022.2063934

Whittlestone, Jess; Arulkumaran, Kai and Crosby, Matthew. "The Societal Implications of Deep Reinforcement Learning" *Journal of Artificial Intelligence Research*, 70 (May 2021), 1003–1030. https://doi.org/10.1613/jair.1.12360