Locating Responsibility in the Future of Human–AI Interactions

I. Introduction

WHETHER we, as end-users of technology, are aware of it or not, our societies are becoming increasingly entangled in a complex network of interactions with Artificial Intelligence (AI) systems. This goes beyond what is often called 'Human-AI collaboration', involving the broader socio-political systems supporting these technologies [1]. Faced with these complex interactions, society grapples with the timeless question: where does responsibility lie for the consequences or results produced by AI systems or applications, whether they are successful or not? Is it with the human operator, AI developer, user, or the AI agent itself? In the case of failure, AI cannot be held accountable, as such software systems are not yet recognized as separate legal entities [2].

This division of liability between humans and AI is often referred to as the "responsibility gap" addressed by a growing body of literature in legal and ethical studies of AI. The responsibility gap is motivated by the central question of: "*To what extent is it possible to hold the manufacturer/programmer /operator of an autonomous, learning automaton responsible for the actions of the machine?*", [3, p. 181]. This concept underscores the difficulty in clearly defining and attributing responsibility to various actors for the actions and decisions made by AI systems. Some researchers, however, argue against the actuality of this gap, viewing responsibility as a dynamic and flexible process capable of effectively encompassing emerging technological entities, including AI systems [4]. They perceive this disparity as less similar to a 'gap' and more resembling a 'shift' in the way responsibility is distributed among actors involved in shaping the human-AI dynamic [5]—a shift of responsibility from users to those who initiated the machine's actions, such as programmers, AI companies, or even to novel agents such as electronic persons specifically created for this purpose [6]. The concept of electronic personhood, first introduced to establish the legal status of autonomous robots, assigns liability to a critical entities, granting specific rights and attributing responsibilities, as well as legal personhood in cases where they act autonomously.

II. Theme

While the debate around the responsibility gap has become more nuanced recently, the literature still lacks a deeper exploration of how responsibility is shared across the network of humans and AI actors [7]. More specifically, we need a framework for understanding how different actors perceive and frame responsibility, and the resulting implications for the development of responsible AI. This is particularly important because, without addressing these fundamental questions, we risk reducing the concept of responsibility to yet another dubious buzzword within AI research and industry [8].

To achieve this, more interdisciplinary and transdisciplinary research is needed [9]. In particular, insights and methods from the social and psychological sciences could provide a valuable new dimension to the current discourse. Such perspectives can contribute by addressing key questions such as: (i) how do users view responsibility in AI systems when things go wrong? and (ii) do they see AI systems as responsible agents or mere tools, or perhaps both, as recent experiments by Longin et al. [10] suggest? To be sure, we need a better understanding of how and why the distribution of responsibility is different in the success and failure of AI systems, and what that tells us about designing and managing AI systems. Other key questions revolve around whether it is desirable to audit AI systems for responsibility, and whether it is appropriate to rely on subjective judgments to determine the extent of responsibility. As shown by Schoenherr and Thomson's experiments in this special issue, the answer to the last question is likely negative. In their experiment, they consider participants as external auditors of responsibility to review vignettes about AI failures and then assign responsibility to the actors involved in the scenario. The authors discovered that the cumulative responsibility rating exceeded 100%. A direct conclusion from this irrational partitioning might be that responsibility might not be divisible in a pure mathematical sense. We cannot determine the extent of each party's responsibility in human-AI interactions simply based on the number of actors involved or our intuitive probabilistic judgments. This simple yet powerful example shows that this genre of research questions is indeed intriguing and exploring them further is critical for the field. Such investigation will not only enrich discussions on the responsibility gap but, more importantly, pave the way for new research avenues in AI research.

Common with other technological changes that have broad effect, there is the option of a response using law, regulation, or self-regulation (for example, by way of codes) [11]. The nature of the response forms one of the themes of this Special Issue. In each of the papers, there is scope for a regulatory or legal response. The challenge here is to decide *who* is responsible for responsible AI. Does responsibility lie with the AI provider or the state? If it is the state, is the response specific to the jurisdiction or is it a coordinated trans-jurisdictional approach using standards bodies such as the IEEE? The prevalence

Digital Object Identifier 10.1109/TTS.2024.3386247

of AI mentioned above leads to calls to "do something." Law makers want to be seen to be responsive to such calls. However, law – especially common law – is not well-suited to responding to dynamic technological change [12], [13]. The alternative is responsive regulation [14]. Here, there is still a risk that the regulation is responsive to the state of technology of a year ago and not appropriate for today. An under-damped regulatory response can create a "whack-a-mole" phenomenon similar to that identified by the Consumer Policy Research Centre (CPRC) [15] in relation to misleading consumers. The worst outcome is that regulatory or legal intervention, designed to foster responsible AI, chills development by stifling innovation [16]. The European General Data Protection Regulation emphasizes "privacy by design" [17] and the U.K. approach to minimizing online harms uses safety by design" [18]. Ensuring AI responsibility *by design* is, potentially, a way to avoid the chilling of development and the associated effect on innovation. This is consistent with a recent approach proposed by the United Nations Educational, Scientific and Cultural Organization staff member in making AI *ethical by design* [19].

III. In This Special Issue

The first paper [A1] in this Special Issue is co-authored by Jordan Schoenherr of Concordia University and Carleton University, and Robert Thomson of the United States Military Academy. The work adopts a novel approach focusing on how an external agent presented with vignettes of automation failures, assigns responsibility, perceives trustworthiness, and explains the successes and failures of AI. Remarkably but perhaps not unexpectedly, the paper reports that the external agent appeared to allocate responsibility not in a wholly rational manner as noted above. Moreover, the study finds that participants with different profiles had contrasting preferences for explanations, such as functional versus mechanistic, which suggests that expandability might not represent a universal attribute of AI.

The second paper [A2] is written by Bijun Wang of Florida Polytechnic University and Onur Asan, Ting Liao, and Mo Mansouri of Stevens Institute of Technology. This work investigates the perceptions and attitudes of chronic patients towards AI-enabled digital healthcare systems. While recognizing the wide-ranging and transformational nature of this technology, this research also identifies variations in patient perceptions. The paper identifies a range of factors that mediate these perceptions and perhaps surprisingly discovers that their level of knowledge of the technology does not appear to have a direct effect.

The third paper [A3] is written by three researchers at the University of South Carolina: Kausik Lakkaraju, Biplav Srivastava, and Marco Valtorta. The work explores a specific type of AI employing causality as the basis for inference relating to sentiments. Notably, the novel approach introduced in this paper works by perturbing inputs of a sentiment analysis system within a controlled causal setting to assess its sensitivity to specific protected attributes. This methodology provides a principled way to compare performance and thus assess alternative black box systems without revealing their internals.

The fourth paper [A4] is written by Ying Xu, Phillipp Terhorst, Kiran Raja, and Marius Pedersen of Norges teknisk-naturvitenskapelige universitet Institutt for datateknologi og informatikk – Gjovik. Part of this work was carried out during the tenure of a European Research Consortium for Informatics and Mathematics (ERCIM) Alain Bensoussan Fellowship Program. This work analyses both the data used to find deepfakes and the services that offer deepfake detection for evidence of demographic and non-demographic bias. The authors report an in-depth investigation of the five most used detection datasets to analyze bias. The results find that both the datasets and the detectors demonstrate biases. The work provides a springboard for further evaluation and mitigation of bias issues in deepfake detection.

The fifth paper [A5] is written by Arvind Upreti and V. Sridhar of the International Institute of Information Technology in India. The work was partially supported by the Machine Intelligence and Robotics Centre (MINRO) at the International Institute of Information Technology, Bangalore. This article investigates the effects of automation on unemployment and wage inequality. It does so by constructing an evolutionary agent-based model of the IT services labor market. The work surprisingly finds that labor market interventions in the form of low-cost retraining support to displaced workers lowers the unemployment rate in the presence of elevated levels of automation. This low-cost retraining is in areas of adjacent work where human-centric skills that complement technology are required. Using these results, the authors propose policy approaches on automation, work policies, and labor welfare.

Appendix: Related Articles

[A1] J. R. Schoenherr and R. Thomson, "When AI fails, who do we blame? Attributing responsibility in human-AI interactions," *IEEE Trans. Technol. Society*, early access, Mar. 1, 2024, doi: 10.1109/TTS.2024.3370095.

[A2] B. Wang, O. Asan, T. Liao, and M. Mansouri, "The future role of clinical artificial intelligence: View of chronic patients," *IEEE Trans. Technol. Society*, early access, Mar. 7, 2024, doi: 10.1109/TTS.2024.3374647.

[A3] K. Lakkaraju, B. Srivastava, and M. Valtorta, "Rating sentiment analysis systems for bias through a causal lens," *IEEE Trans. Technol. Soc.*, early access, Mar. 11, 2024, doi: 10.1109/TTS.2024.3375519.

[A4] Y. Xu, P. Terhörst, M. Pedersen, and K. Raja, "Analyzing fairness in Deepfake detection with massively annotated databases," *IEEE Trans. Technol. Society*, early access, Feb. 16, 2024, doi: 10.1109/TTS.2024.3365421.

[A5] A. Upreti and V. Sridhar, "Assessing the effect of task automation in labor markets: Case of IT services industry," *IEEE Trans. Technol. Society*, early access, Feb. 14, 2024, doi: 10.1109/TTS.2024.3365423.

References

[1] A. Sarkar, "Enough with "human-AI collaboration"," in *Proc. Extended Abstracts Conf. Hum. Fact. Comput. Syst. (CHI)*, 2023, pp. 1–8.

[2] C. Novelli, "Legal personhood for the integration of AI systems in the social context: A study hypothesis," *AI Soc.*, vol. 38, no. 4, pp. 1347–1359, 2023.

[3] A. Matthias, "The responsibility gap: Ascribing responsibility for actions of learning automata," *Ethic. Inf. Technol.*, vol. 6, no. 3, pp. 175–183, 2004.

[4] D. W. Tigard, "There is no techno-responsibility gap," *Philos. Technol.*, vol. 34, no. 3, pp. 589–607, 2021.
[5] M. Simmler, "Responsibility gap or responsibility shift? The attribution of criminal responsibility in human–machine interaction," *Inf. Commun. Soc.*, pp. 1–21, Jul. 2023, doi: 10.1080/1369118X.2023.2239895.
[6] S. M. C. Avila Negri, "Robot as legal person: Electronic personhood in robotics and artificial intelligence," *Front. Robot. AI*, vol. 8, Dec. 2021, Art. no. 789327, doi: 10.3389/frobt.2021.789327.
[7] A. Strasser, "Distributed responsibility in human–machine interactions," *AI Ethic.*, vol. 2, no. 3, pp. 523–532, 2022.
[8] E. Nabavi and C. Browne, "Leverage zones in responsible AI: Towards a systems thinking conceptualization," *Humanit. Soc. Sci. Commun.*, vol. 10, no. 1, pp. 1–9, Mar. 2023.
[9] E. Nabavi, "Why the huge growth in AI spells a big opportunity for transdisciplinary researchers," *Nature*, vol. 429, p. 429, Apr. 2019, doi: 10.1038/d41586-019-01251-1.
[10] L. Longin, B. Bahrami, and O. Deroy, "Intelligence brings responsibility—Even smart AI assistants are held responsible," *Iscience*, vol. 26, no. 8, 2023, Art. no. 107494.
[11] K. W. Abbott and D. Snidal, "The governance triangle: regulatory standards institutions and the shadow of the state," in *Politics of Global Regulation*, W. Mattli and N. Woods, Eds. Princeton, NJ, USA: Princeton Univ. Press, 2008, pp. 44–88.
[12] L. Bennett Moses, "Adapting the law to technological change: A comparison of common law and legislation courts and parliament," *Univ. New South Wales Law J.*, vol. 26, no. 2, pp. 394–417, 2003.
[13] L. Bennett Moses, "Agents of change," *Griffith Law Rev.*, vol. 20, no. 4, pp. 763–794, Jul. 2014.
[14] J. van der Heijden, "Responsive regulation in practice: A review of the international academic literature," SSRN.com. 2020. [Online]. Available: http://dx.doi.org/10.2139/ssrn.3651924
[15] (Consum. Policy Res. Centre, Melbourne, VIC, Australia). *Duped by Design Manipulative Online Design: Dark Patterns in Australia*. (2022). [Online]. Available: https://cprc.org.au/dupedbydesign/
[16] R. Nicholls, "Platforms, privacy and antitrust: Analysing the Australian digital platforms inquiry," in *The Digital Economy: Regulatory, Contractual and Competition Aspects*, P. P. Viscasillas and A. R. Martín-Laborda, Eds. Valencia, Spain: Tirant lo Blanch, 2021, pp. 51–76.
[17] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR)*. Cham, Switzerland: Springer, 2017.
[18] J. Wright and S. Javid (Parliam, London, U.K.), *Online Harms White Paper*. (2019). [Online]. Available: https://doi.org/www.gov.uk/government/publications
[19] G. Ramos, M. Squicciarini, and E. Lamm, "Making AI ethical by design: The UNESCO perspective," *Computer*, vol. 57, no. 2, pp. 33–43, Feb. 2024, doi: 10.1109/MC.2023.3325949.

EHSAN NABAVI
Responsible Innovation Lab
Australian National Centre for the Public Awareness of Science
The Australian National University
Canberra, ACT 2601, Australia
E-mail: Ehsan.Nabavi@anu.edu.au

ROB NICHOLLS
School of Law
University of Technology Sydney
Sydney, NSW 2007, Australia
E-mail: Rob.Nicholls@uts.edu.au

GEORGE ROUSSOS
School of Computing and Mathematical Sciences
Birkbeck College
University of London
WC1E 7HU London, U.K.
E-mail: g.roussos@bbk.ac.uk

Ehsan Nabavi is a Senior Lecturer of Technology and Society with the Responsible Innovation Lab, Australian National Centre for the Public Awareness of Science, The Australian National University. As an Engineer-Sociologist, his research explores the intersection of technology and society. This includes research on responsible modeling and computing and socio-environmental implications of emerging technologies.

Rob Nicholls (Senior Member, IEEE) received the degree in electronics and communications engineering from the University of Birmingham, and the Ph.D. and M.A. degrees from UNSW Sydney. He is a Professional Fellow with UTS Sydney Law. His research interests focus at the intersection of technology and regulation. Before moving to academia, he had a 30-year career, including working for law firms and the Australian competition regulator. He is an Accredited Mediator, and from 2012 to 2020, he was an Australia's Independent Telecommunications Adjudicator.

George Roussos (Member, IEEE) is a Professor of Pervasive Computing with the Birkbeck College, University of London, where he conducts research on digital healthcare and self-sovereignty in decentralized search.