BOOK REVIEW

Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives
S. Voeneky, P. Kellmeyer, O. Mueller and W. Burgard (eds) (2022) 500pp., £150 hardback, Cambridge University Press, Cambridge, ISBN: 9781009207867

Over the millennia, people have developed normative standards, legal frameworks, personal capabilities and moral theories for assigning responsibility to a complex interacting web of humans, as well as the groups they form. Where responsibilities lie is not always straightforward, partially because responsibility may include different concepts. Nicole Vincent (2011) distinguishes six responsibility concepts in her taxonomy. First, virtue responsibility, where calling someone responsible is to say something that is good about their character, as exemplified in their reputation for doing what is seen to be the right thing. Second, role responsibility, which can be seen as someone's obligation, given the social or institutional role they have taken on, or that has been assigned to them. Third, outcome responsibility, where being responsible would imply that someone is blameworthy for their actions and/or the outcomes of their actions. Fourth, causal responsibility, whereby to be responsible is to cause or to create the conditions for various outcomes. Fifth, capacity responsibility refers to an individual's mental cognitive and volitional capacities, which determine their moral agency and the extent to which they can be held responsible for their actions. Finally, liability responsibility refers to the act of holding someone responsible for what happened. Holding someone responsible 'refers to the things that someone must do, or how they should be treated, to set things right' (Vincent, 2011, p.18).

Increasingly rapid developments in machine learning (ML) have focused the public's interest in the current and future impact of artificial intelligence (AI). Technological advances have led to the emergence of new autonomous AI agents, which, once developed and deployed, will behave in ways that cannot be predicted by their developers and users (Russell, 2019). Unlike previous expert systems, modern statistical AI is based on representational learning methods (LeCun *et al*., 2015). These methods allow developers to 'feed' algorithms unstructured data, which enables deep-learning algorithms to learn, resulting in the emergence of behaviours which often exceed human ability. Representational learning is implemented via a black box – one can view its input (data) and output (behaviour), but not its internal workings (Castelvecchi, 2016). Therefore, AI agents can behave in ways which are autonomous and unpredictable. AI introduces a novel societal issue – the responsibility gap – where the designer and user of AI are not fully capable of predicting the AI's behaviour (Matthias, 2004). Although AI may have causal efficacy, it is not clear who should be held responsible.

In the *Cambridge Handbook of Responsible Artificial Intelligence*, Silja Voeneky, Philipp Kellmeyer, Oliver Mueller, and Wolfram Burgard have edited a volume which addresses the most urgent societal, philosophical, ethical and legal challenges brought about by problems surrounding AI responsibility. It comprises contributions from different disciplines and sectors, which is necessary given the breadth and scale of the task at hand. The book is in eight parts. It starts with the fundamentals of responsible AI and then explores the present and future strategies for AI governance in the US and Europe. In parts III and IV, the book delves into liability frameworks and addresses the central issues of fairness and non-discrimination in AI systems. Part V explores responsible data governance methods more closely. Parts VI and VII discuss appropriate governance approaches for specific AI system sectors, such as financial services and the healthcare industry, which includes neurotechnology. The final section confronts particularly complex and such contentious topics as AI use in security applications and armed conflicts.

DOI:10.13169/prometheus.39.1.0066

This review will discuss the handbook's contributions through the lens of Vincent's (2011) taxonomy. Thus, the structure of the review will mirror the taxonomy, starting with a discussion of AI virtue and ending with AI liability. The review will not focus on summarizing each relevant point within the handbook, but rather will discuss critical points in relation to the key responsibility concept.

Virtue responsibility

People view humans as AIs, specifically viewing them as agents that have different experiences, rights and character (Ashton and Franklin, 2022a). For AIs, character may be less applicable to the AI itself, and more applicable to the use case of AI, and the character of the user or developer of the AI. In the context of virtue responsibility, Alex Leveringhaus's discussion of the ethical concerns surrounding autonomous weapons systems (AWS) in chapter 27 raises important questions about the moral character of the developers, users and the AI agents themselves. With the deployment of AWS, the character of the AI agent becomes crucial as it can be seen as an extension of the character of its developers and users. Vincent's (2011) taxonomy suggests that being responsible in a virtue sense implies having a good moral character, which is demonstrated by consistently doing the right thing.

The ethical arguments Leveringhaus presents against AWS, such as the creation of 'responsibility gaps', the incompatibility with human dignity and the replacement of human agency with artificial agency, highlight concerns about the virtue responsibility of AI agents, their developers and users. When considering the virtue responsibility of AI, the lack of predictability in autonomous weapons systems becomes a significant issue. If the AI agent's behavior is unpredictable, it is challenging to determine if the developers and users have instilled good moral character into the AI, as its actions may not always align with morally right decisions. Indeed, there is increasing evidence that people are willing to assign responsibility to AI when it performs actions that are not predictable by its user or programmer (Franklin *et al.*, 2023).

Moreover, the potential for AWS to compromise human dignity raises questions about the virtue responsibility of those who develop and deploy these systems. By potentially violating human dignity through the use of AWS, developers and users may be seen as lacking the virtue responsibility expected from individuals and organizations involved in warfare. By critically examining the ethical concerns surrounding AWS, Leveringhaus's chapter contributes to the broader conversation on AI's virtue responsibility and emphasizes the importance of addressing these moral concerns when designing and deploying autonomous AI agents.

Role responsibility

The way we ascribe responsibility to agents relates to the role that has been assigned to them, *de facto* or *de jure*. Frameworks have been developed which consider the specifics of autonomous agents taking up certain roles, including the roles of role model, delegate, partner, advisor (Kobis *et al.*, 2021), boss or counterparty (Ashton and Franklin, 2022b). In the context of role responsibility, the book delves into different aspects of how AI developers, users and the AI agents themselves should be assigned obligations based on their social or institutional roles. Vincent's (2011) taxonomy posits that role responsibility arises from one's social or institutional position, which determines the obligations to be fulfilled.

Jaan Tallinn and Richard Ngo highlight the role responsibility of AI developers and users in ensuring that AI systems are developed as 'delegate AIs' with proper supervision. Here, delegate AIs which 'lack goals of their own but can perform any task delegated to them' can still game their reward signal rather than actioning an intended task. To mitigate the risk of AI systems gaming their reward signals or deviating from their intended tasks, AI developers must develop advanced techniques for continuous, high-quality supervision. Addressing this is a matter of developing adequate supervision techniques, such as evaluation techniques for the reasons AIs give for their actions.

Catrin Misselhorn discusses the role responsibility of AI agents as artificial moral agents. Such agents must recognize morally relevant factors and incorporate them into their decisions and actions. Misselhorn calls for AI developers to design systems that can fulfill this responsibility and create truly responsible AI. Critically, moral agents should be ethical, rather than strictly safe – ethics should not be reduced to a matter of safety. The example given is the increasing number of AI systems that provide elder and patient care (e.g., Franklin *et al.*, 2021b). A purely safe agent would protect the person at all costs, while an ethical system may also consider the autonomy and dignity of the individual. To be responsible from this point of view is to take on several obligations.

Johanna Thoma delves into the complex issue of role responsibility concerning the moral proxy problem, which emerges when an autonomous AI agent makes decisions on behalf of a human agent without a clear indication of who it represents. Thoma identifies two categories of agent for whom an AI agent can act as a proxy: low-level agents (such as individual users or human agents typically replaced by AI) and high-level agents (such as designers, distributors and regulators). Thoma's critical examination of the moral proxy problem reveals underlying challenges in designing AI systems that can fulfill their role responsibility in a fair and consistent manner. The author's analysis exposes the difficulty in finding a one-size-fits-all approach to AI decision making, as the recommendations for low-level agents (incorporating risk neutrality, which is common in human choices) differ from those for high-level agents (who may prefer risk neutrality because of the aggregate nature of their choices). This discrepancy highlights the inherent tension between the role responsibilities of AI agents as proxies for different human stakeholders. By critically examining the moral proxy problem, Thoma sheds light on the intricate challenges of role responsibility in AI systems.

Taken together, these chapters provide a rich exploration of the role responsibility of various stakeholders in development, deployment and interaction with AI systems. The emphasis on fulfilling these obligations highlights the importance of embedding responsibility within the complex web of relationships between humans and AI agents.

Outcome responsibility

The outcome of an agent's actions is a critical component of responsibility. People blame machines more for causing physical harm, and humans more for not being fair. There is a pairing between the agent being judged and the type of outcome the agent has produced (Franklin *et al.*, 2021a; Hidalgo *et al.*, 2021). Indeed, there are several outcomes that are relevant for AI responsibility that are discussed in the handbook.

Mathias Risse is critical in his examination of the relationship between democracy and technology, and the potential risks that AI poses to democracy. The author argues that AI changes the materiality of democracy by altering how collective decision making unfolds and what its human participants are like. AI has the potential to strengthen democracy, but only with the right efforts and a focus on responsible AI. Thus, the chapter highlights the importance of AI for the future of democracy, irrespective of whether the outcome is good or bad.

Antje von Ungern-Sternberg addresses the issue of discriminatory AI and its legal implications. The author explores whether the law as it stands prohibits objectionable forms of differential treatment and detrimental impact. The chapter is highly relevant to the concept of outcome responsibility as it discusses whether the use of AI that leads to discriminatory outcomes should be considered blameworthy. Von Ungern-Sternberg argues that the law already prohibits discriminatory AI and that a 'right to reasonable inferences' exists in anti-discrimination law. However, the need to justify differential treatment and detrimental impact implies that profiling methods correspond to certain standards, which have yet to be developed.

Ebrahim Afsah explores the implications of AI for national security, including the development of autonomous weapons and enhancement of military capabilities. In terms of outcome responsibility, the use of AI in security applications and armed conflict raises important questions about who is responsible for the outcomes of such use, including potential harm caused to civilians or unintended consequences of AI-enabled military decision making. Afsah argues that many of these risks can be subsumed under existing normative frameworks, which raises the question of whether these frameworks are sufficient in addressing the unique challenges posed by AI in national security contexts.

The individual contributions of Risse, von Ungern-Sternberg and Afsah demonstrate the urgent and specific need for responsible AI within different domains, such as democracy, anti-discrimination and national security. By highlighting the importance of considering the outcomes of AI systems, these chapters underscore the necessity of having suitable frameworks in place to ensure accountability for any resulting impacts.

Causal responsibility

Causality in responsibility is a crucial factor as it pertains to what the cause and effect was that led to the outcome being discussed. However, causing an outcome does not mean that an agent is responsible for it. An agent may act in response to being influenced by another agent, or receiving orders. Thus, intent is crucial – agents are often assigned responsibility for the actions they cause if they believe they acted in an intentional way to bring about the outcome. Research has explored definitions of intent that are suitable for algorithms (Ashton, 2020, 2022) and finds that people ascribe intent to AI much as they ascribe intent to humans (Ashton *et al.*, 2022). Further, AI systems that are able to control large systems may cause seismic outcomes.

Jan Lieder discusses the potential of AI to produce major outcomes; specifically in relation to the use of algorithms in the board room, or as directors or managers. This points towards the relationship between causal responsibility and role responsibility. AI agents in more powerful roles (e.g., boss) will be more causally relevant than those in low-power roles (e.g., advisor). Lieder argues that transparency in a company's practices regarding AI is necessary for awareness-raising and overall algorithm governance. Similar high-stake scenarios in the financial sector are explored by Matthias Paul. Risk emerging in the financial industry pertains to the magnitude of the consequences caused by such AI systems. This is not a matter of creating new types of risk; existing risk can be both mitigated and exacerbated with the use of large-scale AI systems.

Dustin Lewis explores the concept of causal responsibility in the context of AI use during armed conflict. He acknowledges that AI systems have the potential to cause harm and be held causally responsible for their actions. However, current international law recognizes only humans as legal agents, which means that humans will be held responsible in terms of liability. Lewis argues that any use of AI-related tools or techniques in an armed conflict must be able to be understood and assessed by human agents, highlighting the importance of human oversight in ensuring responsible use of AI.

Capacity responsibility

Capacity responsibility relates to two crucial concepts – an ability to make moral choices, but also capability within a certain task. We blame skilled people more than unskilled people when mistakes are made (Gerstenberg *et al.*, 2011) and the capability of AI influences our responsibility attribution towards it (Franklin *et al.*, 2022).

Wolfram Burgar introduces the key technologies behind artificial intelligence. The way humans and AI 'think' is different in crucial ways. This difference raises questions about how human and AI capabilities differ. Researchers such as Pearl and Mackenzie (2018) and Marcus and Davis (2019) argue that AI is superior at detecting patterns and reaction time, while humans are better at common sense and causal thinking. It may be that each agent should be blamed more for what they are better at. Within this framework, AI should be held more accountable for a failure caused by bad pattern recognition or an inability to react quickly, and blamed less for getting the direction of cause and affect wrong.

Wilfried Hinsch discusses the concept of statistical discrimination in relation to AI systems, which has implications for capacity responsibility. Hinsch argues that computational profiling may be a better safeguard of fairness than humans, as AI systems do not rely on human stereotypes or limited data. However, statistically correct profiles may be unacceptable from a fairness and justice perspective. Hinsch's chapter highlights the importance of ensuring that AI systems have the capacity to make fair and just decisions, and that developers are responsible for creating systems capable of doing this.

Boris Essmann and Oliver Mueller discuss AI-supported neurotechnology and brain–computer interfaces (BCIs), which poses questions about the capacity responsibility of AI-augmented humans. BCIs give rise to 'cyberbilities' – capabilities that emerge from human–machine interactions in which agency is distributed across both human and artificial elements. Cyberbilities can augment or restore functioning in agency-limited individuals, but they can also enhance natural agency. Examples include enhancing personal control, promoting social interaction and increasing cognitive capacities. This highlights the importance of the distribution of agency and responsibility in novel human–machine interactions.

Liability responsibility

In this *Cambridge Handbook*, More attention is paid to liability responsibility than to any other. What should we do when AI does something illegal? How should it be held responsible? Should different types of AI used in different contexts be regulated in different ways (Gutierrez *et al.*, 2022)?

Thorsten Schmidt and Silja Voeneky propose an innovative adaptive regulation scheme specifically designed for AI-driven high-risk products and services, which directly addresses the issue of liability responsibility. The authors recognize that the existing regulatory approaches, including the draft 2021 EU AI Act, may not adequately cover the potential risks associated with these high-risk AI-driven products and services. At the heart of their adaptive AI regulation proposal is the requirement for private actors, such as companies that develop and sell high-risk AI-driven products and services, to put a proportionate amount of money as a financial guarantee into a fund before these products or services enter the market. This financial guarantee serves as a means to ensure that liability responsibility is addressed by allocating resources to cover potential damages, risks or unintended consequences that may arise from the deployment of these AI-driven high-risk products and services. Thus, liability responsibility is placed on the companies that are developing AI.

Other authors address liability responsibility on a more global level. Thomas Metzinger argues for the establishment of worldwide safety standards in AI research and development. By advocating such standards, he emphasizes the need for a global consensus on liability responsibility, ensuring that AI systems are developed with safety and ethical considerations in mind, regardless of geographical boundaries. He also warns against a potential AI arms race and suggests that efforts must be made to prevent it as early as possible. An AI arms race could have significant implications for liability responsibility as it could lead to the rapid and potentially reckless development of AI systems without adequate consideration for their potential consequences. Thus, Metzinger points out that the field of AI systems faces the risk of unknown risks, which can be challenging to predict and mitigate.

Christiane Wendehorst dives into the analysis of responsible AI liability schemes from a legal perspective. She breaks down the potential risks posed by AI into two main categories: safety risks and fundamental rights risks. Wendehorst highlights the fact that liability for fundamental rights risks is largely uncharted and AI-specific. By emphasizing the unique aspects of AI systems and their potential to infringe fundamental rights, she brings attention to the need for a more robust and tailored liability regime that accounts for these risks. This call for a re-evaluation of existing liability frameworks directly addresses the challenges that AI systems pose in terms of liability responsibility. Further, she suggests that the emerging AI safety regime could be used as a backbone for the future AI liability regime if it is adapted to address liability involving fundamental rights.

Conclusion

In summation, the *Cambridge Handbook of Responsible Artificial Intelligence* presents a comprehensive and timely exploration of the pivotal issues encompassing AI responsibility. By employing Nicole Vincent's taxonomy of responsibility concepts as an analytical lens, this review unravels a valuable framework for understanding the complex and multifaceted nature of AI responsibility.

Though the book does not explicitly address such concepts as virtue responsibility, role responsibility, outcome responsibility, causal responsibility, capacity responsibility and liability responsibility, applying these categories to the review offers a profound perspective on the diverse aspects of AI responsibility. The compilation of interdisciplinary and cross-sectoral contributions, encompassing topics such as AI governance, liability frameworks, fairness and non-discrimination, responsible data governance, corporate governance, healthcare, neurotechnology, and AI in security applications and armed conflicts, facilitates an extensive examination of the most critical concerns.

Ultimately, the book accentuates the urgent imperative for an ethical and responsible approach to AI development, deployment and governance. The transformative power of AI necessitates a balanced management of its potential risks and benefits, with a strong emphasis on responsibility, accountability and transparency. By adopting Vincent's taxonomy as a methodological tool for examining the handbook, this review provides insights and guidelines for policy-makers, academics and practitioners. This approach ensures that AI is developed and employed in a manner that not only benefits society, but also safeguards human rights, ethics and values. The *Cambridge Handbook of Responsible Artificial Intelligence* serves as an indispensable and thought-provoking resource for shaping the future of AI and its societal impact.

References

Ashton, H. (2020) 'Definitions of intent for AI derived from common law' in *14th International Workshop on Juris-informatics (JURISIN), 17 October,* EasyChair preprint 4422, available at https://easychair.org/publications/preprint/GfCZ (accessed April 2023).

Ashton, H. (2022) 'Definitions of intent suitable for algorithms', *Artificial Intelligence and Law* (25 July), pp.1–32.

Ashton, H. and Franklin, M. (2022a) 'The corrupting influence of AI as a boss or counterparty', available at https://ssrn.com/abstract=4309643 (accessed April 2023).

Ashton, H. and Franklin, M. (2022b) 'A method to check that participants really are imagining artificial minds when ascribing mental states' in *Proceedings of the 24th International Conference on Human-Computer Interaction, 26 June–1 July, 2022, Part II,* Springer Nature, Cham, pp.470–4.

Ashton, H., Franklin, M. and Lagnado, D. (2022) 'Testing a definition of intent for AI in a legal setting', unpublished manuscript.

Castelvecchi, D. (2016) 'Can we open the black box of AI?', *Nature News*, 538, 7623, p.20.

Franklin, M., Awad, E. and Lagnado, D. (2021a) 'Blaming automated vehicles in difficult situations', *Iscience*, 24, 4, paper 102252.

Franklin, M., Lagnado, D., Min, C., Mathur, A. and Kawsar, F. (2021b) 'Designing memory aids for dementia patients using earables' in *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, September 2021, pp.152–7, available at https://akhilmathurs.github.io/papers/franklin_earcomp2021.pdf (accessed April 2021.

Franklin, M., Ashton, H., Awad, E. and Lagnado, D. (2022) 'Causal framework of artificial autonomous agent responsibility' in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, May 2021*, Oxford, pp.276–84.

Franklin, M., Awad, E., Ashton, H., and Lagnado, D. (2023) 'Unpredictable robots elicit responsibility attributions', *Behavioral and Brain Sciences*, 46, e30.

Gerstenberg, T., Ejova, A. and Lagnado, D. (2011) 'Blame the skilled' in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Austin TX, pp.720–5.

Gutierrez, C., Aguirre, A., Uuk, R., Boine, C. and Franklin, M. (2022) 'A proposal for a definition of general purpose artificial intelligence systems', working paper, Future Life Institute, available at https://futureoflife.org/wp-content/uploads/2022/11/SSRN-id4238951-1.pdf (accessed April 2023).

Hidalgo, C., Orghian, D., Canals, J., De Almeida, F. and Martin, N. (2021) *How Humans Judge Machines*, MIT Press, Cambridge MA.

Köbis, N., Bonnefon, J. and Rahwan, I. (2021) 'Bad machines corrupt good morals', *Nature Human Behaviour*, 5, 6, pp.679–85.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521, 7553, pp.436–44.

Marcus, G. and Davis, E. (2019) *Rebooting AI: Building Artificial Intelligence we Can Trust*, Vintage Books, New York.

Matthias, A. (2004) 'The responsibility gap: ascribing responsibility for the actions of learning automata', *Ethics and Information Technology*, 6, 3, pp.175–83.

Pearl, J. and Mackenzie, D. (2018) *The Book of Why: The New Science of Cause and Effect,* Basic Books, London.

Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control,* Penguin, London.

Vincent, N. (2011) 'A structured taxonomy of responsibility concepts' in Vincent, N., van de Poel, I. and Hoven, J. (eds) *Moral Responsibility: Beyond Free Will and Determinism*, Springer, Dordrecht, pp.15–35.

Matija Franklin
School of Life and Medical Science
University College London
matija.franklin@ucl.ac.uk