ROBONOMICS
The Journal of the Automated Economy
https://journal.robonomics.science

Book review of Yampolskiy, R.V. (2024). *AI: Unexplainable, Unpredictable, Uncontrollable*. Chapman and Hall/CRC. 264 pp. ISBN: 978-1-003-44026-0.

Reviewed by

Katerina Volchek [1] ✉

Type: Book review

Citation: Volchek, K. (2024). Book review of Yampolskiy, R.V. (2024). AI: Unexplainable, Unpredictable, Uncontrollable. Chapman and Hall/CRC. 264 pp. ISBN: 978-1-003-44026-0. *ROBONOMICS: The Journal of the Automated Economy*, 5, 64

[1] Deggendorf Institute of Technology, Germany; Email: katerina.volchek@th-deg.de
✉ Corresponding author

Publication history
Received: 21/04/2024; Revised: 05/05/2024; *Accepted*: 05/05/2024; *Published online*: 17/05/2024; *Volume date*: 31/12/2024

Artificial intelligence (AI) has already transformed multiple industries in recent decades. On the one hand, it serves as a key resource for innovations (Holmström, 2022). AI's capabilities for data analytics and service automation will become a determinant of business competitiveness. Special-purpose algorithms like cancer detection, crime prediction, and car driving automation have been affordable for large enterprises with proactive R&D developments (Nassif et al., 2022; Kyrkou et al., 2020). At the same time, various AI-driven applications, such as face recognition, text, photo, and video generation, have already become everyday commodities. The capabilities of different types of AI speed up its adoption by businesses and individuals.

On the other hand, it has been acknowledged that AI's capabilities can lead to such a degree of transformation that the planned tasks fail or go out of control. While the full potential of AI has not yet been realised, multiple issues related to incorrect results have been observed. Thus, AI has been observed to produce gender and racial bias (Dai et al., 2022; Gupta et al., 2022). Generative AI has been acknowledged as "hallucinating" not only the content but also its sources (Carvalho & Ivanov, 2024). The calls for developing new principles of "Explainable AI" and even a new philosophy of "Digital Humanism" have been posted to optimise the use of AI, while preventing possible damage from it (Angelov et al., 2021; Werthner et al., 2022). Elon Musk predicts AI to overcome human intelligence by 2029 (Hammond, 2024). However, the essence and the challenges of AI remain unknown and misunderstood.

The book „AI: Unexplainable, Unpredictable, Uncontrollable" is a timely response to the developing field of artificial intelligence. It provides an overview of the phenomenon of AI and a straightforward explanation of AI controllability issues. The book highlights AI's risks and dangers regardless of its objectives. While creating a bit of negative connotation, it offers a comprehensive summary of the potentially negative consequences of AI implementation.

First, the book explains the need to control AI as a potentially independent actor in contemporary society. While the advantages of using AI are nearly unlimited, the authors warn about an almost infinite number of AI failures due to the system's Unexpectability, Unpredictability and Non-Verifiability. Being integral properties of an AI system, they make the AI-based decisions erroneous. The book, therefore, starts by discussing two approaches: controlling AI capabilities to minimise possible damage and controlling AI design to prevent damage. The author calls for both approaches to be considered when designing and governing a safer AI.

The book proceeds with a detailed explanation of AI properties that determine AI safety and the positive outcome of its application. Thus, Unpredictability, Unexplainability & Incomprehensibility, Unverifiability and Unownability are discussed in separate chapters, providing the reader with an introduction to each concept and an explanation of its reasons and consequences. This enables a discussion on the interference of human participation in the above-mentioned phenomena and the possible consequences.

The next logical part of the book concludes with a broad discussion on AI controllability as a reflection of a human and/or technological capacity to manage AI. This section summarises that regardless of the intentions, the development of AI inevitably leads to unplanned control issues due to its above-named properties. According to the author, full control over AI is impossible by definition. The impossibility of control equally restricts the full implementation of both beneficial and malicious AI-driven projects. Prevention of the potentially negative consequences of AI requires special design principles and integrated safety rules.

Further on, the book contains chapters dedicated to safety and governance principles for AI. Thus, the chapters "Pathways to Danger" and "Accidents" highlight the dangers that AI can cause. The author explains that the negative effects of AI proliferation are very complex and multifaceted. They include accidental AI errors, an

unintended industry transformation due to the substitution of a manual workforce with AI-enabled robotics, and the purposeful design of a dangerous or unethical code.

Last but not least, the book extends the discussion to the concepts of AI personhood and AI consciousness, opening a question of the independence of AI decision-making and the ownership and responsibility for these decisions. While AI is a human-made algorithm, it already exceeds the capacity of a human brain. While it doesn't have a real "consciousness," i.e., it could not fully "feel" the value of made decisions, AI is already "smarter" than any human, i.e., it can process much more information. Inevitably, AI started dominating multiple domains of human life. The book, therefore, claims that instead of being sceptical of the entire AI field it is vital to focus on solutions to design safe AI such as AI-driven personal universes with properties designed for each person, to address some of the AI safety issues and prevent possible damage.

Although the book's structure is not very transparent, it allows readers to develop a comprehensive picture of the objective AI challenges alongside possible human-developed biases and motives. The book is suitable for researchers and practitioners in the fields of Software development and Mathematics. While it doesn't provide any clear answers, the book creates a background for a safer, more manageable and more ethical design of AI algorithms.

This book would also be beneficial for business managers and strategists who explore the pros and cons of incorporating AI in their organisations' business models. It aligns with the ideas of "Digital Humanism" (Werthner et al., 2022) in its discussion about the role of AI and Humans in the current society. The discussed dimensions of AI raise awareness of the possible challenges an AI-driven business might face. As a result, the book contributes to minimising the risks of negative impacts of AI, enabling its safer application and to raising a call for more advanced AI design.

References:

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

Carvalho, I., & Ivanov, S. (2024). ChatGPT for tourism: applications, benefits and risks. *Tourism Review,* 79(2), 290-303. https://doi.org/10.1108/TR-02-2023-0088

Dai, B., Xu, Z., Li, H., Wang, B., Cai, J., & Liu, X. (2022). Racial Bias Can Confuse AI for Genomic Studies. *Oncologie,* 24(1), 113-130. https://doi.org/10.32604/oncologie.2022.020259

Gupta, M., Parra, C. M., & Dennehy, D. (2022). Questioning racial and gender bias in AI-based recommendations: Do espoused national cultural values matter? *Information Systems Frontiers*, *24*(5), 1465-1481. https://doi.org/10.1007/s10796-021-10156-2

Hammond, G. (2024). *Elon Musk predicts AI will overtake human intelligence next year*. Financial Times. Retrieved 21 April 2024 from https://www.ft.com/content/027b133f-f7e3-459d-95bf-8afd815ae23d

Holmström, J. (2022). From AI to digital transformation: The AI readiness framework. *Business Horizons*, *65*(3), 329-339. https://doi.org/10.1016/j.bushor.2021.03.006

Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022). Breast cancer detection using artificial intelligence techniques: A systematic literature review. Artificial Intelligence in Medicine, 127, 102276. https://doi.org/10.1016/j.artmed.2022.102276

Kyrkou, C., Papachristodoulou, A., Kloukiniotis, A., Papandreou, A., Lalos, A., Moustakas, K., & Theocharides, T. (2020, July). Towards artificial-intelligence-based cybersecurity for robustifying automated driving systems against camera sensor attacks. In 2020 IEEE computer society annual symposium on VLSI (ISVLSI) (pp. 476-481). IEEE. https://doi.ieeecomputersociety.org/10.1109/ISVLSI49217.2020.00-11

Werthner, H., Prem, E., Lee, E. A., & Ghezzi, C. (2022). *Perspectives on digital humanism*. Springer Nature.