Information Research - Vol. 30 No. 3 (2025)

Handbook on the ethics of artificial intelligence

Review of Gunkel, David J. (Ed.). (2024). Handbook on the ethics of artificial intelligence. Cheltenham, UK: Edward Elgar. x, 326 p. ISBN: 978-1-80392-671-1

DOI: https://doi.org/10.47989/ir30360315

Although "artificial intelligence" (or, more commonly, "AI") is the buzzword of the day, nothing akin to human intelligence actually exists in the world of computer-based "thinking machines" (except, perhaps, in the world of mathematics–see Nunez, 2025). The responses of generative AI systems, for example, which appear to be capable of carrying on an intelligent conversation with a human being, are based on the statistical manipulation of texts, with billions of parameters involved. They cannot think and have no capacity for moral behaviour other than whatever guidelines are programmed into the system. Even when "instructed" not to lie, they will generate fictitious data, illustrating an inability to understand the relationship between lying and "hallucinating". In a sense, therefore, the title of this book is misleading: artificial intelligence (at least as we know it today) has no ethics–rather, the title ought to have been, *Handbook on the ethics of the use of artificial intelligence.*

There are twenty-one chapters in the book and the Introduction tells us that they are divided into five sections, covering the initial issue of why an ethical position is needed, through rights and responsibilities in the use of AI, to politics and alternative philosophies of AI use. It's a little odd that these are not signalled in the contents list–the reader will have to mark the divisions.

The wide geographical range of contributors is to be welcomed, since it provides views influenced by a wide range of circumstances. Thus, the twenty-one chapters have thirty-one authors, seven (plus the Editor) from the USA; seven from the UK; Netherlands, five; Germany and Australia, three; two from North Macedonia, and one each from France, Hungary, Singapore and Thailand. However, the Editor has not only achieved geographical representation, he has also ensured that alternative views of ethics, for example, from a feminist and from a Buddhist perspective.

A thorough review of all chapters would result in a very lengthy review, so I have selected those chapters that interested me, or that I thought ought to have attention drawn because of what I see as their significance.

Appropriately, the first section begins with a chapter on the nature of ethics and why it is necessary to have an ethics of AI, by Sven Nyholm of Munich's Ludwig-Maximilians University. Nyholm distinguishes between a narrow conception of ethics, which is concerned with what is, and is not, permissible, and the broader conception, which includes the narrow, but is extended to cover such concepts as what it is to lead a good life. He suggests that a narrow conception of the ethics of AI would focus on negative issues such as aspects of AI that lead to harm or injustice. The broader conception has more positive concerns such as how AI might contribute to a more meaningful life. For example, if AI systems enable us to remove tedious, time-consuming aspects of our work, and allow us to engage with meaningful issues to a greater extent, we may experience greater job satisfaction.

Nyholm does not present conclusions about what kind of ethical framework might be needed for AI, but, rather, sets us thinking about what kind of issues need to be addressed in ethical terms.

In the final chapter of the first section, Rebecca Johnson of the University of Sydney, discusses "What are responsible AI researchers really arguing about?", identifying two communities: the "AI-ethics" community, which is "concerned with the immediate repercussions of AI on individuals, societies, and vulnerable populations"; and the "AI-safety" community, which is "more focused on existential threats to humanity..." Johnson does not attempt to resolve the differences between the two groups, but points to the different epistemological bases of their arguments. She suggests that the AI-safety community has a basis in functionalism, while the AI-ethics community is based on constructivism. She goes on to argue that alternative epistemological positions are needed for AI ethics and suggests that 4E cognition (embodied, embedded, enacted, and extended), evolved in cognitive science is an appropriate basis for AI ethics, since it "emphasises that real 'understanding' emerges from active interactions rather than just passive information absorption". Enactivism is somewhat similar in that in emphasises agency and autonomy and the agent's interaction with the environment.

The second section of the book deals with the issue of responsibility: given that AI systems are tools used by humans, ultimate responsibility for the actions achieved by AI lies with the human. Since AI systems (so far, at least) are not moral beings, they cannot be held responsible for the consequences of their use. From this section I've chosen just one chapter, the last: "From ethics to law: why, when, and how to regulate AI" by Simon Chesterman, of the National University of Singapore. I chose this chapter because the issue of how to regulate the use of generative AI systems in universities is highly relevant at this moment.

Chesterman outlines the range of regulatory possibilities, from instruments such as codes of conduct to legislative frameworks, and suggests the most appropriate for different circumstance. He draws attention to the limitations of voluntary measures, noting that legislative oversight becomes essential as the potential impacts on society eventuate. The discussion refers to case studies that highlight the global unevenness of regulatory developments.

Section three deals with moral values and rights, and I found the chapter on anthropomorphism by Eleanor Sandry (Curtin University, Australia) interesting, since there is a definite tendency to treat generative AI systems as-if they were human. We use a conversational style in the prompts use, ask for something in polite language, and say "Thank you", as if we were talking to a human being. Sandry argues (with reference to other scholars) that "anthropomorphism is the only way humans can encounter and attempt to communicate with non-human others", and concludes that the process does not need to treat the other as-if human, but simply as-if partially human, thereby remaining "aware of the otherness of AI technologies while also supporting human relations with them".

The chapters of section four deal with power and politics–dealing with issues ranging from the impact on the environment, to the impact on disadvantaged person of the inequalities of access. In "Disabling AI: biases and values embedded in artificial intelligence", Damien Williams (University of North Carolina at Charlotte), argues that AI systems of various kinds have negative biases and values embedded in them. We know this, of course, from reports on facial recognition systems leading to false identification of people, to the failures of self-drive cars (Iwersen and Verfürden, 2025). Williams deals specifically with racist biases that result from the already biased databases of images that the systems are trained on. He also deals with gender issues, class and capitalist issues, and biases that affect persons with disabilities. He suggests that things can change, "by ensuring that the perspectives and lived experiences of marginalized people are heeded in conversations about the design and implementation of algorithmic applications..." That is a big ask, and, I suspect, one that will be difficult to implement.

My choice from section five, is "Buddhism and the ethics of artificial intelligence" by Soraj Hongladarom (Chulalongkorn University, Bangkok, Thailand). Hongladarom is the author of a book

on the application of Buddhist ethics to AI and robotics (Hongladarom, 2020), and in this chapter he summarises the main argument of that book and addresses the criticisms that have been raised as a result. The essential point that the author makes is that a central tenet of Buddhism is that the world is one of *dukkha*–suffering or unsatisfactory states of being, and that actions that aim to decrease dukkha are good, and those that lead to an increase of dukkha are bad. Hence, any AI system that leads to an increase in any unsatisfactory state should be avoided, and AI that leads to a decrease, and, ultimately to *nirvana*–or release from unsatisfactory states of being, is to be approved. Thus, while use of generative AI to prepare an essay may initially feel good to a student, subsequently their dukkha will be increased as a result of any punitive action taken against them. Similarly, a self-drive car may free the driver from the routine act of driving and free time for other activity while in the car, if the AI system in the vehicle creates circumstances through which the person is harmed or even killed, such a system cannot be an ethical AI system. The author's arguments that a Buddhist ethics of AI provide a different perspective from either deontological or consequentialist ethics seems to be entirely justified.

This is a timely collection of papers, but it is in no sense a "handbook" – a point made about another collection from the same source. Do not expect to be offered a set of guidelines for determining whether or not the use of AI systems is ethical or unethical. That is not the purpose of the book. Rather, it offers a wide variety of views on the nature of ethics and their application to AI, revealing the complexity of the subject. The reader must choose for themselves how to engage ethically with AI.

I was a little surprised that the authors of the chapter refer so infrequently to the established ethical philosophies of Aristotle, Kant, and Bentham, and the more recent ethicists such as Nussbaum, Korsgaard, and O'Neill. The lack of engagement with the ethical theories of these authors is puzzling.

Prof. T.D. Wilson
Professor Emeritus, University of Borås
July, 2025

References

Hongladarom, S. (2020). *The ethics of AI and robotics: a Buddhist viewpoint*. Lanham, MD: Lexington Books.

Iwersen, S. & Verfürden, M. (2025, July 5).'The vehicle suddenly accelerated with our baby in it': the terrifying truth about why Tesla's cars keep crashing. *The Guardian.* https://tinyurl.com/2u8mndra

Nuñez, M. (2025, July 21). Google Deep Mind makes AI history with gold medal win at world's toughest math competition. *VentureBeat*. https://tinyurl.com/ycxef4rt

Problems in facial recognition. (2025, July 12). *Geeks for geeks.* https://www.geeksforgeeks.org/blogs/problems-in-facial-recognition/