Philosophy in Review

Toby Walsh, "Machines Behaving Badly: The Morality of AI"

Thomas Klikauer

Volume 43, Number 4, November 2023

URI: https://id.erudit.org/iderudit/1108428ar
DOI: https://doi.org/10.7202/1108428ar

See table of contents

Publisher(s)

University of Victoria

ISSN

1206-5269 (print)
1920-8936 (digital)

Explore this journal

Cite this review

Klikauer, T. (2023). Review of [Toby Walsh, "Machines Behaving Badly: The Morality of AI"]. *Philosophy in Review*, *43*(4), 43–45. https://doi.org/10.7202/1108428ar

Toby Walsh. *Machines Behaving Badly: The Morality of AI.* La Trobe University Press & Black Inc. Press. 2022. 275pp; AUD $34.99 (Paperback 9781760643423).

Just as the title of the book—*The Morality of AI*—indicates, Toby Walsh's recent work is about the morality of artificial intelligence (AI). Walsh's book is not about mathematics, the writing of algorithms, and it is not even about the computing logic of artificial intelligence. Instead, it is about the point where AI meets morality.

This is followed by 'people' and 'companies'. While the fourth chapter is about 'autonomy'. This is then followed by 'human v. machines'. Eventually, his work arrives at 'ethical rules' discussing two important issues for AI: 'fairness' and 'privacy'. The final two chapters are on 'the planet'", i.e., global warming and sustainability, and 'the way ahead.' The book ends with a short 'epilogue.' While the book and this review are about artificial intelligence, both are not written by the now (in)famous ChatGPT algorithmic essay writing website, *OpenAI.com/blog/ChatGPT*.

Walsh starts his chapter on AI by making clear, in 'reality, artificial intelligence is [not a] conscious robot … [w]e cannot yet build machines that match the intelligence of a two-year-old … program computers … do narrow, focused tasks' (1). Like a coffee maker—a machine—AI does not have morality. Still, every time someone asks Siri—Apple's virtual assistant—a question, this person uses AI. At the same time, AI is based on 'machine-learning algorithms that [can, for example, also] predict which criminals will reoffend' (2)—which, for Walsh is problematic.

Walsh correctly argues that there is a 'a common misconception … that AI is a single thing. Just like our intelligence is a collection of different skills, AI today is a collection of different technologies' (3). Surprisingly and after decades of research into AI, Walsh freely and correctly admits, 'we have made almost no progress on building more general intelligence that can tackle a wide range of problems' (3).

Despite this, AI does have an impact on society. The author emphasises that 'we walked straight into [a] political minefield in 2016, first with the Brexit referendum in the United Kingdom and then with the election of Donald Trump in the United States. Machines are now routinely treating humans mechanically and controlling populations politically' (11). Worse, 'Facebook can be used to change public opinion in any political campaign' (11).

Yet, when writing an algorithm for artificial intelligence, those who write these scripted codes are often, white men—Walsh calls it the 'sea of white dudes' problem' (15). And worse, the sea of white dudes is also overrepresented in the venture capital that turns AI into profits. These venture capital companies can be divided into 'three roughly equally sized parts: Silicon Valley the rest of America, and the rest of the world' (21).

Today, we see, for example, that 'Apple's turnover is more than the GDP of Portugal' (39). Worse, 'most Big Tech companies are sitting on large cash mountains. It is estimated that US companies have over $1 trillion of profits waiting in offshore accounts'—a nice word for *tax havens*, an immoral and rather shady tax-reduction scheme. Besides the semi-criminal side of big tech corporations, Walsh is not worried about super-intelligent machines that bypass human beings. He notes that many of those 'working in AI, are not greatly worried about … super-intelligent machines

[still] the philosopher Nick Bostrom … fears that super-intelligence poses an existential threat to humanity's continuing existence.… Suppose we want to eliminate cancer. "Easy," a super-intelligence might decide: "I simply need to get rid of all hosts of cancer." And so, it would set about killing every living thing!' (43).

In short, clever corporate and highly manipulative marketing disabled many customers' autonomy by making people addicted to tobacco (first) and online platforms (later). Yet, for AI, the autonomy of human beings remains a very serious issue. Walsh argues, 'autonomy … is an entirely novel problem. We've never had machines before that could make decisions independently of their human masters' (61). On this, he alludes to 'the trolley problem' (77) or 'trolleyology' (78), as he calls it.

This brings Walsh to MIT's moral machine which itself has moral problems. MIT's www.moralmachine.net basically asks you to vote on 'moral issues' (83). Yet, Walsh argues that voting on morality is a dicey issue saying, 'we humans often say one thing but do another. We might say that we want to lose weight, but we might still eat a plate full of delicious cream doughnuts' (83). Yet, there is also a second problem. Unlike real elections, MIT's 'moral machine [is] not demographically balanced' (8). It is used by young college-educated men—Walsh's sea of white dudes.

Yet what we might call life, AI calls 'living systems' [which is a system] that maintain some sort of equilibrium, have a life cycle, undergo metabolism, can grow and adapt to their environment, can respond to stimuli, and can reproduce, and evolve' (104). 'In fact, there's a branch of AI called genetic programming that has evolution at its core' (104;) —never mind genetic programming's Knapsack problem.

Apart from all this, Walsh is convinced that, '[t]here's no need—indeed, no place—for free will' (106). One of the great things about Walsh's book is that it consistently reminds readers that 'computers are deterministic machines that simply follow the instructions in their code' (106). Next to this fact, morality might come with the following point which might not really concern us in the very short-term: 'Once machines are conscious, we may have ethical obligations to them in how we treat them. For instance, can we now turn them off?' (107).

Furthermore, one thing that distinguishes us from AI are emotions. On that, AI people believes that 'there are six basic emotions: anger, disgust, fear, happiness, sadness, and surprise' (108)—thankfully love does not seem to be an emotion for Walsh's sea of white dudes. Perhaps that is why we might be better off to 'imagine AI—as an alien (rather than artificial) intelligence' (113).

Things are getting a bit more serious with 'Google Translator … translating [for example] he is pregnant' correctly into the correct but rather nonsensical German of, *er ist schwanger* (113). Perhaps such nonsense is simply because AI is 'unable to experience pain and suffering, then it might follow that, no matter how intelligent they become, they are not in need of any rights. We can treat them like any other machine. Toasters have no rights' (115).

AI and morality are highly important in the realm of 'medicine' (140) where morality and AI remain two of the most problematic aspects. This leads Walsh to the morality of fairness. Walsh uses the example of the UK's Ofqual where 'students from poor state schools were more likely to have

their grades marked down than students from rich private schools' (149).

Even graver things have happened around policing. Walsh notes that ''Predicting future crime [while using] historical data will then only perpetuate past biases.… Those who use AI to learn from history are doomed to repeat it.' In fact, it's worse than repetition. We may construct dangerous 'feedback loops in which we magnify the biases of the past' (151). Particularly on the issue of fairness in policing, there is also the danger of 'a common mistake in machine learning where we confuse correlation with causation' (156). On the morality of fairness, Walsh says, 'there are now at least 21 different mathematical definitions of fairness in use by the machine-learning community' (158). Perhaps fairness is not a mathematical definition.

Yet fairness also plays into something as simple as speech recognition. Walsh lists speech-recognition systems such as those, for example, developed by Amazon, etc., and notes that, 'all five systems performed significantly worse for black speakers than for white speakers. The average word error rate across the five systems was 35 per cent for black speakers, compared with just 19 per cent for white speakers' (168). In other words, 'AI systems will often reflect the biases of the society in which they are constructed' (171) as well as the bias of those who create algorithms, i.e., a sea of white dudes.

Meanwhile, 'training [the] enormous model [of ChatGPT-3] is estimated to have produced 85,000 kilograms of CO_2. To put this in perspective, this is the same amount produced by four people flying a round trip from London to Sydney in business class [and exactly as much as] people in a row of eight economy-class seats [use] for the same trip [and in] practice, many data centres run on renewable energy. Google Cloud claims to have zero net carbon emissions' (212). At the same time, 'transportation is responsible for around one-quarter of global CO_2 emissions. AI can be a great asset in reducing these emissions' (217).

Walsh concludes his insightful book by saying, 'It should be clear by now that we cannot today build moral machines, machines that capture our human values and that can be held accountable for their decisions.… Machines will always and only ever be machines. They do not have our moral compass.… If we can't build moral machines, it follows that we should limit the decisions that we hand over to machines' (222) – a rather logical and inevitable conclusion.

Finally, and while not trying to be a moral philosopher, Walsh has produced an exquisite book that delivers fascinating insights into morality and artificial intelligence; those who create AI – the sea of white dudes; and into the corporations that use AI, often to the detriment of society. In the end, Walsh has produced a thoughtful and insightful book on the *Morality of Artificial Intelligence*.

Thomas Klikauer, Western Sydney University