Volume **4** (2) 2022 https://doi.org/10.5204/lthj.2677

Book Review

Toby Walsh (2022) *Machines Behaving Badly: The Morality of AI*. Collingwood: La Trobe University Press and Black Inc Books

Nicholas Korpela
Queensland University of Technology, Australia

ISBN: 9781760643423

There has been recent interest and speculation in the media as to whether the artificial intelligence (AI) system developed by Google, in the form of the Language Model for Dialogue Applications (LaMDA), is sentient.[1] *Machines Behaving Badly: The Morality of AI* is a timely contribution at a time when AI is under much greater scrutiny due to the significant investments by companies like Facebook, Google and other venture capitalists, who are seeking to advance engineering breakthroughs. According to *The AI Index 2021 Annual Report* from Stanford University, the total global investments in AI reached USD67.9 billion in 2020.[2] The pandemic also had a negligible effect on AI investments, and in fact, participation at conferences increased due to the capacity of researchers to attend virtually.[3] With all this in mind, Walsh delivers a comprehensive overview of ethics applied to AI and provides the reader with 10 chapters of an education on the subject that is well presented and infused with clever witticisms. From the outset, Walsh challenges the Hollywood conceptions of AI, such as *The Terminator* and *Blade Runner*, remarking, 'why is it that AI is always trying to kill us? In reality, [AI] is not one of these conscious robots. We cannot yet build machines that match the intelligence of a two-year-old'.[4]

Walsh explains AI is more subtle than these Hollywood depictions as they already exist in Tesla vehicles, machine learning algorithms and speech recognition software such as Siri.[5] The claim that Google's LaMDA project may prove to be sentient was made by Blame Lemoine, a senior engineer who was suspended by Google after releasing their chat logs to the public.[6] This story raises an important question: How would data scientists reasonably converge on defining and measuring the sentience of AI? Would the Turing test be the only viable convergence? Machine learning AI is ultimately a reflection of its human creators; it is created in our image, data and digital footprint. People who have the particular set of skills required to build AI programs form a very small part of the world's population, and Walsh explains that AI programmers only represent a relatively small group.[7] There is also more evidence suggesting those with neurodiversity are attracted to science, technology and mathematics fields, which is not a new theory, but it is certainly understudied as it relates to scientists working within the AI

[1] Vallance, "Google Engineer."
[2] Zhang, "AI Index," 93.
[3] Zhang, "AI Index," 2, 83.
[4] Walsh, Machines Behaving Badly, 9.
[5] Walsh, Machines Behaving Badly, 9.
[6] Walsh, Machines Behaving Badly, 9.
[7] Walsh, Machines Behaving Badly, 9.

field.[8] In this way, neurodivergent individuals may envision or think about AI technology and its usage differently from the general population. Knowing this technology is in the hands of very few people does pose significant questions about how AI is deployed and its implications for society as a whole.

Walsh further reminds readers that most venture capitalists live in Silicon Valley, which also has a meaningful effect on how AI is developed.[9] This was an important consideration for Walsh, as such a group becomes reflective of the dominant values and culture existing within this critically important industry. When pairing venture capitalists with the relatively small number of AI technology specialists in the world, it leaves little room to wonder how techno-libertarianism surfaced as a movement.[10] In a nutshell, techno-libertarianism refers to the idea that technology should not be restrained by regulations or any kind of control and that technology should only be limited by the ingenuity of its innovators.[11] Walsh critically discusses the importance of understanding the people and culture behind the development of AI applications and refers to them mostly as 'the sea of dudes' problem.[12] Therefore, it is logical that the creators of the technology, and their identities and backgrounds, inform the ethical rules, or lack thereof, associated with its use.[13] The author also cites government use of AI technology to be problematic, which exacerbates pre-existing concerns over mass surveillance and racial prejudices.[14]

Walsh highlights that one of the more easily exploited applications of AI concerns digital platforms exploiting users' personal data without their consent and the use of facial recognition and biometric technology.[15] The author believes facial recognition software 'may one day be less biased than humans'.[16] Even if this is technically true, it is optimistic to suggest that humans will consistently manage this technology ethically, including the right to privacy. The People's Republic of China is an example of a government that unambiguously weaponises a technological armamentarium against its citizens through the deployment of biometrics that is coupled with a social credit system.[17] It seems human history has exposed a propensity for authoritarianism to emerge in every generation somewhere in the world, and AI technology has only intensified this power. For this reason, AI algorithms applied within the context of our legal system are also contentious.[18] As a case study, Walsh examined the Correctional Offender Management Profiling for Alternative Sanctions tool, which has been used to predict who is at risk of reoffending and resulted in racial minority groups receiving false positives.[19] It should be clear that there are considerable implications to the rule of law when a machine with an unreliable success rate is making decisions of significant consequence.[20]

While this raises concerns on its own, the extreme end of the spectrum would be the development and deployment of lethal autonomous weapon systems.[21] This issue has been raised by the United Nations and many different parts of the world that want a ban on lethal autonomous weapons.[22] Conversely, countries like Turkey, Russia, the United States and Australia are investing heavily in autonomous weapons for the purposes of military deployment, and these investments are not without their scrutiny.[23] It is almost inconceivable that Russia, a state that is currently under investigation by the Office of the Prosecutor of the International Criminal Court for war crimes in Ukraine, is developing an autonomous nuclear-powered and nuclear-tipped submarine referred to as *Poseidon*.[24] Walsh does not answer whether government officials operating without restraint is worse than an algorithm that could make those decisions for itself, but does accept the history of wars have demonstrated human beings are quite capable of acquiescing in genocide and crimes against humanity.[25] Notwithstanding, the idea of creating an efficient killing machine with an ethical code is not only an oxymoron but equally problematic. This is because even if it is theoretically possible to embed the laws of armed conflict into its programming, it could still be vulnerable to cyber-attacks,

[8] Wei, "Science, Technology," 2.
[9] Walsh, Machines Behaving Badly, 20.
[10] Walsh, Machines Behaving Badly, 22.
[11] Walsh, Machines Behaving Badly, 28.
[12] Walsh, Machines Behaving Badly, 20.
[13] Walsh, Machines Behaving Badly.
[14] Walsh, Machines Behaving Badly, 41.
[15] Walsh, Machines Behaving Badly, 112, 147.
[16] Walsh, Machines Behaving Badly, 147.
[17] Liang, "Constructing a Data-Driven Society", 415, 420; Tsai, "Hobbling Big Brother", 10.
[18] Walsh, Machines Behaving Badly, 119.
[19] Walsh, Machines Behaving Badly.
[20] Walsh, Machines Behaving Badly, 120.
[21] Walsh, Machines Behaving Badly, 53.
[22] Walsh, Machines Behaving Badly.
[23] Walsh, Machines Behaving Badly, 54.
[24] Walsh, Machines Behaving Badly, 55; International Criminal Court, "Statement of ICC Prosecutor".
[25] Walsh, Machines Behaving Badly, 55.

electromagnetic interference, or instrument malfunction.[26] Nevertheless, it seems the limitations of the United Nations, and geopolitics more generally, is an inadequate deterrence as global powers have already commenced a lethal autonomous weapon arms race for this century.[27]

Ironically, the most well-known ethical rules of AI are based on science fiction, as Walsh reflects further on Isaac Asimov's 'three laws of robotics'.[28] The three laws primarily serve to protect human beings from harm, even if ordered to do so, and include the robot prioritising human lives above itself.[29] Walsh believes these laws are too simplistic and do not account for the complexity of rules that could include utilitarian ethics.[30] It is obvious that killer robots being deployed against humans is one measure of harm, but Walsh provides more pervasive examples of how algorithms can influence capital markets or decisions of health insurance providers.[31] Delegating decisions to machines, as Walsh suggests, does not necessarily result in fairness or avoid 'cognitive biases' and has become a feature rather than a bug.[32]

Towards the end of *Machines Behaving Badly*, Walsh dedicates the ninth chapter to the current climate emergency.[33] It is argued that with the assistance of big tech companies, AI technology could enable oil companies like Chevron Corporation to extract fossil fuels with more efficiency.[34] On this point Walsh provides readers with a stark statistic: '100 companies are responsible for 71 per cent of global emissions'.[35] Moreover, companies like ExxonMobil, BP, Chevron Corporation and Shell are reaping an unprecedented windfall of profits in the midst of the current inflationary prices of petroleum, diesel and gas.[36] The ongoing harm to our collective environment and their objectionable profiteering during a time of war and inflation provide no good reasons why big tech should continue to assist big oil.[37] Indeed, this is a subtle example of how AI contributes to climate change, and it is not hyperbolic to suggest such applications could have considerable consequences for human civilisation in the long term. While this is a distinct contrast from *The Terminator* feature films, the outcomes of AI used to harvest and pollute natural resources would be eerily similar in terms of its outcome and destruction of life on earth.

In the final chapter, Walsh acknowledges that humans are currently incapable of developing moral machines and, therefore, must not be delegated 'high-stakes decisions'.[38] In terms of trusting AI in the future, Walsh considers, first, the need for AI to encompass a range of desirable characteristics such as 'explainability, auditability, robustness, correctness, fairness, respect for privacy, and transparency',[39] second, that a tightening of regulation is needed, and third, that it is vital the public be educated on how they should utilise emerging technology.[40] These arguments are convincing because this education could include providing the community with a better understanding of how AI is integrated into the digital platforms they use, minimising risks associated with the exploitation of their data, and why their informed consent is both necessary and important.[41]

Machines Behaving Badly was a timely book to review, given its contemporaneity and contribution to the ongoing discussion of the ethical dilemmas concerning AI. I would recommend this book to those who are interested in knowing more about the current myths, research, and future challenges of AI and robotics. There is already a pervasiveness of AI in modern society, and what was considered illusory back when Turing first envisaged AI has essentially come to pass. It was also encouraging to read Walsh's feminist critiques of the culture within big tech, as well as his concerns in relation to them partnering up with big oil, and that this could accelerate the world's climate emergency.[42] While the author raises the many ways in which AI can go wrong and the need for further technical, regulatory, and educational advancements going forward, he provides several different counterpoints to balance this perspective.

[26] Walsh, Machines Behaving Badly.
[27] Walsh, Machines Behaving Badly, 75.
[28] Walsh, Machines Behaving Badly, 74.
[29] Walsh, Machines Behaving Badly, 74.
[30] Walsh, Machines Behaving Badly.
[31] Walsh, Machines Behaving Badly, 101.
[32] Walsh, Machines Behaving Badly, 138.
[33] Walsh, Machines Behaving Badly, 161.
[34] Walsh, Machines Behaving Badly, 165.
[35] Walsh, Machines Behaving Badly.
[36] Winck, "2 wild charts."
[37] Walsh, Machines Behaving Badly, 165.
[38] Walsh, Machines Behaving Badly, 170.
[39] Walsh, Machines Behaving Badly, 171–73.
[40] Walsh, Machines Behaving Badly, 177, 178, 182.
[41] Walsh, Machines Behaving Badly.
[42] Walsh, Machines Behaving Badly, 119, 143.

There is still hope that engineers, scientists, governments, and people more generally, can collaborate with each other and use AI for the overall benefit of society. However, therein lies the limitations of *Machines Behaving Badly*, and that is that no one can predict how humanity will ultimately overcome the many ethical challenges with respect to an automated society. This is even more uncertain in an era of global competition, and when countries such as the United States and China are at risk of falling into the 'Thucydides Trap', which is an observed pattern throughout history of when a rising power threatens the hegemony of the current world order there is a high probability of war.[43] Vladimir Putin has also expressed an opinion that when it comes to AI: "whoever becomes the leader in this sphere will become ruler of the world", which may prove to be a concerning prospect for the future of humanity, especially when it comes to using AI superiority for the purposes of war, power and control.[44]

Bibliography

Allison, Graham. "The Thucydides Trap." *Foreign Policy*, June 8, 2017. https://foreignpolicy.com/2017/06/09/the-thucydides-trap/

International Criminal Court. "Statement of ICC Prosecutor, Karim A.A. Khan QC, on the Situation in Ukraine: 'I have decided to proceed with opening an investigation,' " February 28, 2022 https://www.icc-cpi.int/news/statement-icc-prosecutor-karim-aa-khan-qc-situation-ukraine-i-have-decided-proceed-opening

Liang, Fan, Vishnupriya Das, Nadiya Kostyuk, and Muzzammil M. Hussain. "Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure". *Policy and Internet* 10, no 4 (2018), 415-453. https://doi.org/10.1002/poi3.183

Maggio, Edoardo. "Putin Believes That Whatever Country has the Best AI will be 'The Ruler of the World.' " *Business Insider*, 4 September 4, 2017. https://www.businessinsider.com/putin-believes-country-with-best-ai-ruler-of-the-world-2017-9

Tsai, Wen-Hsuan, Hsin-Hsien Wang, and Ruihua Lin, 'Hobbling Big Brother: Top-Level Design and Local Discretion in China's Social Credit System', *The China Journal* 86, (2021). https://doi.org/10.1086/714492

Vallance, Chris. "Google Engineer Says LaMDA AI System May Have Its Own Feelings." *BBC*, June 14, 2022. https://www.bbc.com/news/technology-61784011

Walsh, Toby. *Machines Behaving Badly: The Morality of AI*. Collingwood: La Trobe University Press and Black Inc Books, 2022.

Wei, Xin, Jennifer W Yu, Paul Shattuck, Mary McCracken and Jose Blackorby. "Science, Technology, Engineering, and Mathematics (STEM) Participation Among College Students with an Autism Spectrum Disorder." *Journal of Autism and Developmental Disorders* 43, no 7 (2013): 1539–46. https://doi.org/10.1007%2Fs10803-012-1700-z

Winck, Ben and Madison Hoff. "2 Wild Charts Show How Big Oil Profits are Skyrocketing as Prices Pump at the Pump Rise." *Business Insider*, May 18, 2022. https://www.businessinsider.com/gas-prices-oil-company-profits-skyrocketing-energy-sector-earnings-charts-2022-5

Zhang, Daniel, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Depp Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark and Raymond Perrault. *The AI Index 2021 Annual Report*. (AI Index Steering Committee, Human-Centered AI Institute, Stanford University, 2021). https://arxiv.org/ftp/arxiv/papers/2103/2103.06312.pdf

[43] Allison, "The Thucydides Trap."
[44] Maggio, "Putin Believes That Whatever Country has the Best AI."