**Ethics and Artificial Intelligence**

Dr Mark Ryan, Wageningen Economic Research, Wageningen University & Research, [mark.ryan@wur.nl](mailto:mark.ryan@wur.nl)

**Synonyms:**

Artificial intelligence – AI

**Introduction**

We have been encapsulated by images of humanoid robots, often portrayed in science fiction movies and TV series, that look, speak, and act, like us. This human mimicry is what is often what is understood as ‘artificial intelligence’ (AI) in popular culture. It is typically the most inviting, interesting, and visual, picture of AI, that allows us to understand the potential of this technology. These types of AI robots often take on the role as our assistants, colleagues, friends, and even lovers. Conversely, media has also portrayed AI through dystopian images, such as Skynet in the Terminator franchise, whereby, these machines turn against us and try to either enslave or eradicate us altogether. AI is conveyed as either leading us to utopia or being our downfall.

While these depictions are still in the world of science fiction, AI has started to receive increased attention over the past decade as something that will propel the economy and humanity like no technology before it. There is certainly a lot of hype around how AI is the future for business, and can solve many challenges we face today, throughout a wide array of domains (healthcare, transportation, insurance, agriculture, and administration, to name a few). However, what is meant by ‘AI’ is sometimes vague, which makes it difficult to identify the ethical issues that we need to consider.

**What is AI?**

AI is typically understood as the performance of intelligence by digital technology, to a degree independent from human control. Of course, what classifies as acts of intelligence and to what degree they must be independent from human control is open for debate. How to distinguish between advanced (non-AI) technological solutions and AI solutions is often challenging, and many people group a wide array of digital applications under the ‘AI’ banner. For example, recently, it was reported that 40% of ‘AI start-ups’ do not even use any form of AI (Vincent 2019). There is such a huge buzz around AI that everyone wants to be seen using it.

AI is intelligence that is different from natural intelligence, as it is constructed and created through digital technologies. These advanced intelligent systems have the capability to perform complex tasks normally associated with humans, ‘such as image recognition (vision), speech recognition (hearing), and natural language generation (speaking). AI is artificial mimicry of tasks and functions that would otherwise require human intelligence’ (Ryan, 2020).

AI takes the form of computer programmes and software to process, predict, and prescribe actions. This type of AI is being used to determine whether one should be granted bank loans or insurance pay-outs, identifying areas that police should patrol, criminals’ likelihood to reoffend, provide prognosis for cancer patients, detect fraud, and provide tailored recommendations for end-users.

AI can also be embodied within robotic machinery, taking a humanoid form, or in the form of cars, drones, farming machinery, medical devices, and so forth. However, not all robots have AI, there are also ‘dumb robots’ that are completely under the control of a human agent or do extremely restrictive tasks (e.g., bomb disposal units, robots in manufacturing plants, etc.). Robots that have AI often create their own unique challenges, which will be shown throughout this chapter.

**What is AI Ethics?**

A subdiscipline has emerged around AI ethics, which is comprised of a wide array of individuals: computer scientists, ethicists, cognitive scientists, roboticists, legal professionals, economists, sociologists, gender, and race theorists. This has led to a very interesting branch of research, addressing issues surrounding the development and use of AI. This chapter will give a very brief snapshot of some of the most pertinent ethical concerns. Many of the issues in the Big Data Ethics chapter in this collection are often applicable to AI ethics, because of the data that these technologies retrieve, store, and use, so will not be duplicated here.[[1]](#footnote-1)

While data-related issues are not new or unique to AI, AI does hold the potential to dramatically retrieve and analyse data that would not be possible otherwise. Data is being used in unique and transformative ways, such as the use of facial recognition to identify individuals from photos or CCTV; AI robots to retrieve live video-feed about the patient(s) that it is monitoring; or in new ways, such as self-driving cars collecting an abundance of data about our surroundings, how we drive, and passengers in the car. There is the potential to infringe on individuals’ privacy, restriction of resources, or at worst, the creation of a surveillance society.

**AI and Discrimination**

AI is being used to make predictions about future events, but also, in many contexts, it provides prescriptions of actions that should be taken. With the inclusion of AI in the decision-making process, it has raised several problematic concerns related to unfair bias, discrimination, and justice. There is the possibility that unfair biases arise from the *data that is inputted* into the AI model, the *types and ranges of data used*, the code within the *algorithm* itself, or biases in how the recommendations (*output data*) from AI are interpreted.

For example, self-driving vehicles’ computer vision has recently been shown to have a much lower accuracy to detect people of colour, putting them at a much higher risk (Cuthbertson 2019). Banks, insurance companies, and HR, are using AI to identify risk individuals will not pay back loans, are committing insurance fraud, or suitable candidate based on their personal data. Much of these recommendations are based on data, such as the individual’s postcode, names, gender, and the university they attended. In other words, AI has been making assumptions based on data that is closely linked to one’s ethnicity, gender, class, and background.

Even scarier is the idea that AI can be used to ‘predict’ our behaviour, personality, moods, and even likelihood to commit (or re-commit, for reference see COMPAS in the US (e.g., see Rahman 2020)) crimes (spoiler: it cannot). In the latter example, AI is making these predictions based on racially-biased datasets and on racial feature similarities. This modern form of phrenology is hugely problematic and is a form of pseudoscience that keeps being used, despite being discredited time-and-time again.

AI is also being used by police to predict potential crime hotspots, based on historical data, using predictive algorithms. However, there is the threat that AI will identify impoverished areas, leading to over-policing and feelings within the community of constantly feeling monitored, and/or it will cause other areas to be neglected and under-policed. This would have a self-perpetuating and reinforcing effect on crime, as the data confirms that these were indeed crime hotspots, thus advocating for their increased patrolling.

**AI and Manipulation**

There is also the potential that different forms of AI will manipulate our behaviour. For example, the use of targeted advertising, using advanced AI methods, to identify and customise ads and lure us into buy their products. There is a much stronger capacity for companies to specifically target individuals with the use of AI. AI has also recently been used to manipulate individuals in voting, by creating, circulating, or reinforcing, misleading content; targeting those who may be swayed one way (see the 2016 US Presidential Elections and the United Kingdom’s Leave Brexit Poll).

There are also many forms of AI that are specifically built to deceive and manipulate, such as ‘deep fakes’, which create fake video, audio, or image content that looks or sounds exactly as though it were real. Deep fakes manipulate individuals into believing that certain actions occurred, individuals behaved in a certain way, and altogether, manipulating our perceptions of reality. There have already been problems with deep fake ‘revenge porn’ and there is also a concern about their potential to incite social unrest, violence, and war. This use of AI poses grave threats of disinformation, which thus impacts our informed decision-making and actions.

**AI and Responsibility**

It is unclear who is responsible for the actions of AI. As this technology will be learning, behaving, and interacting with humans, relatively independently, there is concern about responsibility allocation when things go wrong. Who should be held responsible when self-driving vehicles’ malfunction and cause a crash (see the 2016 Uber death in Arizona); AI programs that say racist or harmful things (see Microsoft’s Tay); or it underscores school grades, stopping people from getting into the university they were hoping for (see the UK’s 2020 algorithmic grading debacle)?

As AI acts autonomously from human action, it is difficult to classify it as simply an instrument for human action. AI can independently make important decisions about our lives. Traditional instruments can be misused by the user, or they can malfunction because of a manufacturing fault; however, the autonomous actions caused by AI may materialise outside of these traditional parameters. Furthermore, the design and use of AI often brings a very large range of stakeholders together, making it quite difficult to pinpoint the stage of development that led to the harmful event.

The type of tasks that AI should be given is also quite controversial. The level of responsibility of AI, its potential risk, and its capacity to fulfil these activities can be difficult to identify. Some have proposed that greater understandability, transparency, and ‘explainable AI’ will allow us to trace the causes behind an AI’s action, thus, making it easier to allocate responsibility. However, there is often a trade-off between developing AI that is more effective, but may have less understandability about how it arrived at those decisions (some may be completely unknown, otherwise, known as ‘black-box actions’), and a form of AI that is more understandable, but perhaps less effective.

**AI and Autonomy**

There is much debate about what level of autonomy AI should possess, with one of the main concerns surrounding a future where there is difficulty to differentiate AI agency from human agency (Allen et al. 2006; and Crnkovic and Çürüklü 2012). AI will have an increased level of independence from human agency, so some claim that they need to have a built-in ethical response to real-world situations, becoming “artificial moral agents” themselves (Crnkovic and Çürüklü 2012). Utilitarianism, deontology, and virtue ethics have been proposed as ethical frameworks that could be programmed into AI to guide their behaviour (Wallach and Allen 2009). Others propose implementing AI with bottom-up approaches to ethics or building-up moral competencies through learning itself (Anderson and Anderson 2015). Crnkovic and Çürüklü (2012) add that the use of AI will vary, and so too will the level of autonomy in specific contexts and applications. Many of the same issues that arise with all these ethical frameworks will emerge when implemented within AI, but creating many new, challenging issues, with one of the most challenging: should they even have this kind of capacity or agency in the first place?

**AI and Moral Agency**

One of the concerns around AI is whether it should be given the capacity to become a moral agent. However, another one is how do we know that AI is acting as a moral agent and not simply emulating how a moral agent should act. Some have suggested developing a ‘Moral Turing Test’, whereby, an interviewer asks AI a series of questions, to see if they can distinguish the machine from a human’s moral reasoning. There is also the ‘Comparative Moral Turing Test’, where the interviewer is asked to compare the answers of a human and AI to see if they can identify if the robot is giving a less moral answer of the two (Allen, et al. 2000). However, both tests are flawed as they still do not answer whether the AI understands what they are saying or if it is just a carefully, and intelligently, reiterated phrase that materialises from their programming. Establishing a method that can determine the moral agency of an AI is challenging and is an ongoing area of research within the field.

**AI and Robot Rights**

Closely aligned to analysing the moral agency of AI is how we behave and interact with AI, and what type of moral considerability should it receive. At what level should we begin granting AI with specific rights and acknowledging its moral considerability? Typically, this question gets asked in the context of AI robots, but some still speculate about our behaviour towards non-embodied AI (take for instance, the fictional HAL 9000 from 2001: A Space Odyssey). The field is divided between those who believe we can and/or should, while others state that we cannot and/or should not, attach rights to AI robots.

Kate Darling (2016) claims that we should grant AI robots with rights because it feels wrong to harm them, regardless of their capacities. Some propose that AI will have a higher form of intelligence in the future that will require us to grant them rights, so we should pre-emptively do this before they demand them (Asaro 2006). In contrast, Joanna Bryson (2010) proposes that we *could* design machines with this level of intelligence, but that we *should not*, as it is morally undesirable. We should not create AI that necessitate the granting of human-like rights to these machines. Whereas others propose that we will never be able to design a machine, nor should we, that would necessitate the granting of rights, because they are still simply artefacts of human creation (Johnson 2006). While Gunkel (2018) believes that we should grant AI robots moral consideration regardless of their ontology, because this is how we act in the real-world; namely, we behave ethically towards others without knowing their inner capacity or ontological development.

**AI’s Effects on Human Relationships**

The widespread use of AI may have a dramatic effect on human-human interactions. If AI becomes advanced enough that people are unsure if they are interacting with a human or AI, it may cause us to be very apprehensive when interacting with others, for the fear that we are interacting with a robot. This is an issue that is commonly referred to the ‘uncanny value’, where people report being comfortable interacting with a human-like AI robots up to a certain point. When that likeness is indistinguishable, it becomes undesirable and unnerving. Lifelike AI may also impact the way that we behave towards other human beings. The objectification and instrumentalization of robots may cause us to behave in a similar way towards other human beings.

Using AI in certain contexts may also replace many valuable human interactions. Replacing humans with AI in health and care settings, holds the potential to dehumanise, objectify, and instrumentalise those patients. Respect of the individual patient and their sense of dignity would be harmed if there is a sense that these ‘burdens’ should be taken care of by AI robots, instead. Our interactions with the elderly, vulnerable, and persons with disabilities, may be directed towards AI, creating, or exacerbating, feelings of isolation, loneliness, and debasing their humanity.

**AI and Unemployment**

There is also a huge concern around employment when AI is set to replace many human jobs, as they will be able to do them more efficiently, for less money, and without much of the ‘hassle’ of employing humans (insurance, healthcare, trade unions, etc.). However, what will become of the individuals who used to work in these jobs, and will it negatively transform some sectors entirely, as a result? There is a tension in the field, with some proposing that AI will allow us to have more free-time, reducing the length of work-weeks, and will replace many of the ‘dirty, dull, and dangerous’ jobs within society. Individuals can retrain, upskill, and altogether, do safer, cleaner, and more interesting work.

However, there is also scepticism, with fears that the capitalistic society in which we live, will simply force individuals in (often) low-paid and low-skilled jobs out of jobs, with no opportunities for future employment. While technologization and its effect on employment is nothing new, the sheer range and abundance of jobs that will be under threat by AI-replacements is unparalleled, so requires careful attention to ensure that this transition is desirable, ethical, and balanced for society.

**Cross-references**

Artificial Intelligence and Ethical Journalism; Artificial Intelligence and Teaching Values in Science; Big Data Ethics; Engineering Ethics; Ethics of Ambient Assisted Living Technologies; Technology Ethics: Origins, Paradigms and Implications for Business and Society

**References**

Allen, Colin, Gary Varner, and Jason Zinser. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12. Taylor & Francis: 251–261.

Allen, Colin, Wendell Wallach, and Iva Smit. 2006. Why machine ethics? *IEEE Intelligent Systems* 21. IEEE: 12–17.

Anderson, Michael, and Susan Leigh Anderson. 2015. Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot: An International Journal*. Emerald Group Publishing Limited.

Asaro, Peter M. 2006. What should we want from a robot ethic? *The International Review of Information Ethics* 6: 9–16.

Bryson, Joanna J. 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* 8. John Benjamins Amsterdam: 63–74.

Cuthbertson, Anthony. 2019. Self-driving cars more likely to drive into black people, study claims. The Independent. Accessed on 19th February 2021. https://www.independent.co.uk/life-style/gadgets-and-tech/news/self-driving-car-crash-racial-bias-black-people-study-a8810031.html

Crnkovic, Gordana Dodig, and Baran Çürüklü. 2012. Robots: ethical by design. *Ethics and Information Technology* 14. Springer: 61–71.

Darling, Kate. 2016. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *Robot law*. Edward Elgar Publishing.

Gunkel, David J. 2018. The other question: Can and should robots have rights? *Ethics and Information Technology* 20. Springer: 87–99.

Johnson, Deborah G. 2006. Computer systems: Moral entities but not moral agents. *Ethics and information technology* 8. Springer: 195–204.

Rahman, Farhan. 2020. COMPAS Case Study: Fairness of a Machine Learning Model. *Towards Data Science*. https://towardsdatascience.com/compas-case-study-fairness-of-a-machine-learning-model-f0f804108751

Ryan, Mark. 2020. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics* 26. Springer: 2749–2767.

Vincent, James. 2019. Forty percent of ‘AI startups’ in Europe don’t actually use AI, claims report. *The Verve*. Accessed on 19th February 2021. https://www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report

1. Many of the challenges faced in the ethics of Big Data correlate with AI ethics. This is because vast streams of data are being retrieved, stored, cleaned, processed, and used, by AI applications. Many of the issues discussed in the Big Data Ethics chapter are equally applicable here, therefore. [↑](#footnote-ref-1)