

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/349467117>

Ethics of Artificial Intelligence

Article · February 2021

CITATIONS

36

READS

18,852

2 authors, including:



Sven Nyholm

Ludwig-Maximilians-Universität in Munich

110 PUBLICATIONS 2,437 CITATIONS

SEE PROFILE

Ethics of Artificial Intelligence

This article provides a comprehensive overview of the main ethical issues related to the impact of Artificial Intelligence (AI) on human society. AI is the use of machines to do things that would normally require human intelligence. In many areas of human life, AI has rapidly and significantly affected human society and the ways we interact with each other. It will continue to do so. Along the way, AI has presented substantial ethical and socio-political challenges that call for a thorough philosophical and ethical analysis. Its social impact should be studied so as to avoid any negative repercussions. AI systems are becoming more and more autonomous, apparently rational, and intelligent. This comprehensive development gives rise to numerous issues. In addition to the potential harm and impact of AI technologies on our privacy, other concerns include their moral and legal status (including moral and legal rights), their possible moral agency and patienthood, and issues related to their possible personhood and even dignity. It is common, however, to distinguish the following issues as of utmost significance with respect to AI and its relation to human society, according to three different time periods: (1) short-term (early 21st century): autonomous systems (transportation, weapons), machine bias in law, privacy and surveillance, the black box problem and AI decision-making; (2) mid-term (from the 2040s to the end of the century): AI governance, confirming the moral and legal status of intelligent machines (artificial moral agents), human-machine interaction, mass automation; (3) long-term (starting with the 2100s): technological singularity, mass unemployment, space colonisation.

Table of Contents

1. The Relevance of AI for Ethics
 - a. What is AI?
 - b. Its Ethical Relevance
2. Main Debates
 - a. Machine Ethics
 - i. Bottom-up Approaches: Casuistry
 - ii. Top-down Approaches: The MoralDM Approach
 - iii. Mixed Approaches: The Hybrid Approach
 - b. Autonomous Systems
 - c. Machine Bias

- d. The Problem of Opacity
- e. Machine Consciousness
- f. The Moral Status of Artificial Intelligent Machines
 - i. The Autonomy Approach
 - ii. The Indirect Duties Approach
 - iii. The Relational Approach
 - iv. The Upshot
- g. Singularity and Value Alignment
- h. Other Debates
 - i. AI as a form of Moral Enhancement or a Moral Advisor
 - ii. AI and the Future of Work
 - iii. AI and the Future of Personal Relationships
 - iv. AI and the Concern About Human ‘Enfeeblement’
 - v. Anthropomorphism
3. Ethical Guidelines for AI
4. Conclusion
5. References and Further Reading

1. The Relevance of AI for Ethics

This section discusses why AI is of utmost importance for our systems of ethics and morality, given the increasing human-machine interaction.

a. What is AI?

AI may mean several different things and it is defined in many different ways. When Alan Turing introduced the so-called Turing test (which he called an ‘imitation game’) in his famous 1950 essay about whether machines can think, the term ‘artificial intelligence’ had not yet been introduced. Turing considered whether machines can think, and suggested that it would be clearer to replace that question with the question of whether it might be possible to build machines that could imitate humans so convincingly that people would find it difficult to tell whether, for example, a written message comes from a computer or from a human (Turing 1950).

The term ‘AI’ was coined in 1955 by a group of researchers—John McCarthy, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon—who organised a famous two-month summer workshop at Dartmouth College on the ‘Study of Artificial Intelligence’ in 1956. This event is widely recognised as the very beginning of the study of AI. The organisers described the workshop as follows:

We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. (Proposal 1955: 2)

Another, later scholarly definition describes AI as:

the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience. (Copeland 2020)

In the early twenty-first century, the ultimate goal of many computer specialists and engineers has been to build a robust AI system which would not differ from human intelligence in any aspect other than its machine origin. Whether this is at all possible has been a matter of lively debate for several decades. The prominent American philosopher John Searle (1980) introduced the so-called Chinese room argument to contend that strong or general AI (AGI)—that is, building AI systems which could deal with many different and complex tasks that require human-like intelligence—is in principle impossible. In doing so, he sparked a long-standing general debate on the possibility of AGI. Current AI systems are narrowly focused (that is, weak AI) and can only solve one particular task, such as playing chess or the Chinese game of Go. Searle's general thesis was that no matter how complex and sophisticated a machine is, it will nonetheless have no 'consciousness' or 'mind', which is a prerequisite for the ability to *understand*, in contrast to the capability to *compute* (see section 2.e.).

Searle's argument has been critically evaluated against the counterclaims of functionalism and computationalism. It is generally argued that intelligence does not require a particular substratum, such as carbon-based beings, but that it will also evolve in silicon-based environments, if the system is complex enough (for example, Chalmers 1996, chapter 9).

In the early years of the twenty-first century, many researchers working on AI development associated AI primarily with different forms of the so-called machine learning—that is, technologies that identify patterns in data. Simpler forms of such systems are said to engage in 'supervised learning'—which nonetheless still requires considerable human input and supervision—

but the aim of many researchers, perhaps most prominently Yann LeCun, had been set to develop the so-called self-supervised learning systems. These days, some researchers began to discuss AI in a way that seems to equate the concept with machine learning. This article, however, uses the term ‘AI’ in a wider sense that includes—but is not limited to—machine learning technologies.

b. Its Ethical Relevance

The major ethical challenges for human societies AI poses are presented well in the excellent introductions by Vincent Müller (2020), Mark Coeckelbergh (2020), Janina Loh (2019), Catrin Misselhorn (2018) and David Gunkel (2012). Regardless of the possibility of construing AGI, autonomous AI systems already raise substantial ethical issues: for example, the machine bias in law, making hiring decisions by means of smart algorithms, racist and sexist chatbots, or non-gender-neutral language translations (see section 2.c.). The very idea of a machine ‘imitating’ human intelligence—which is one common definition of AI—gives rise to worries about deception, especially if the AI is built into robots designed to look or act like human beings (Boden et al. 2017; Nyholm and Frank 2019). Moreover, Rosalind Picard rightly claims that ‘the greater the freedom of a machine, the more it will need moral standards’ (1997: 19). This substantiates the claim that all interactions between AI systems and human beings necessarily entail an *ethical dimension*, for example, in the context of autonomous transportation (see section 2.d.).

The idea of implementing ethics within a machine is one of the main research goals in the field of machine ethics (for example, Lin et al. 2012; Anderson and Anderson 2011; Wallach and Allen 2009). More and more responsibility has been shifted from human beings to autonomous AI systems which are able to work much faster than human beings without taking any breaks and with no need for constant supervision, as illustrated by the excellent performance of many systems (once they have successfully passed the debugging phase).

It has been suggested that humanity’s future existence may depend on the implementation of solid moral standards in AI systems, given the possibility that these systems may, at some point, either match or supersede human capabilities (see section 2.g.). This point in time was called ‘technological singularity’ by Vernon Vinge in 1983 (see also: Vinge 1993; Kurzweil 2005; Chalmers 2010). The famous playwright Karl Čapek (1920), the renowned astrophysicist Stephen Hawking and the influential philosopher Nick Bostrom (2016, 2018) have all warned about the possible dangers of technological singularity should intelligent machines turn against their creators, that is, human beings. Therefore, according to Nick Bostrom, it is of utmost importance to build friendly AI (see the alignment problem, discussed in section 2.g.).

In conclusion, the implementation of ethics is crucial for AI systems for multiple reasons: to provide safety guidelines that can prevent existential risks for humanity, to solve any issues re-

lated to bias, to build friendly AI systems that will adopt our ethical standards, and to help humanity flourish.

2. Main Debates

The following debates are of utmost significance in the context of AI and ethics. They are not the only important debates in the field, but they provide a good overview of topics that will likely remain of great importance for many decades (for a similar list, see Müller 2020).

a. Machine Ethics

Susan Anderson, a pioneer of machine ethics, defines the goal of machine ethics as:

to create a machine that follows an ideal ethical principle or set of principles in guiding its behaviour; in other words, it is guided by this principle, or these principles, in the decisions it makes about possible courses of action it could take. We can say, more simply, that this involves “adding an ethical dimension” to the machine. (2011: 22)

In addition, the study of machine ethics examines issues regarding the moral status of intelligent machines and asks whether they should be entitled to moral and legal rights (Gordon 2020a, 2020b; Richardson 2019; Gunkel and Bryson 2014; Gunkel 2012; Anderson and Anderson 2011; Wallach and Allen 2010). In general, machine ethics is an interdisciplinary sub-discipline of the ethics of technology, which is in turn a discipline within applied ethics. The ethics of technology also contains the sub-disciplines of robot ethics (see, for example, Lin et al. 2011, 2017; Gunkel 2018; Nyholm 2020), which is concerned with questions of how human beings design, construct and use robots; and computer ethics (for example, Johnson 1985/2009; Johnson and Nissenbaum 1995; Himma and Tavani 2008), which is concerned with commercial behaviour involving computers and information (for example, data security, privacy issues).

The first ethical code for AI systems was introduced by the famed science fiction writer Isaac Asimov, who presented his Three Laws of Robotics in *Runaround* (Asimov 1942). These three were later supplemented by a fourth law, called the Zeroth Law of Robotics, in *Robots and Empire* (Asimov 1986). The four laws are as follows:

1. A robot may not injure a human being or, through inaction, allow a human being to be harmed;
2. A robot must obey the orders given it by human beings except where such orders would conflict with the first law;
3. A robot must protect its own existence as long as such protection does not conflict with the first or second law;

4. A robot may not harm humanity or, by inaction, allow humanity to suffer harm.

Asimov's four laws have played a major role in machine ethics for many decades and have been widely discussed by experts. The standard view regarding the four laws is that they are important but insufficient to deal with all the complexities related to moral machines. This seems to be a fair evaluation, since Asimov never claimed that his laws could cope with all issues. If that was really the case, then Asimov would perhaps not have written his fascinating stories about problems caused partly by the four laws.

The early years of the twenty-first century saw the proposal of numerous approaches to implementing ethics within machines, to provide AI systems with ethical principles that the machines could use in making moral decisions (Gordon 2020a). We can distinguish at least three types of approaches: bottom-up, top-down, and mixed. An example of each type is provided below (see also Gordon 2020a: 147).

i. Bottom-up Approaches: Casuistry

Guarini's (2006) system is an example of a bottom-up approach. It uses a neural network which bases its ethical decisions on a learning process in which the neural network is presented with known correct answers to ethical dilemmas. After the initial learning process, the system is supposed to be able to solve new ethical dilemmas on its own. However, Guarini's system generates problems concerning the reclassification of cases, caused by the lack of adequate reflection and exact representation of the situation. Guarini himself admits that casuistry alone is insufficient for machine ethics.

ii. Top-down Approaches: The MoralDM Approach

The system conceived by Dehghani et al. (2011) combines two main ethical theories, utilitarianism and deontology, along with analogical reasoning. Utilitarian reasoning applies until 'sacred values' are concerned, at which point the system operates in a deontological mode and becomes less sensitive to the utility of actions and consequences. To align the system with human moral decisions, Dehghani et al. evaluate it against psychological studies of how the majority of human beings decide particular cases.

The MoralDM approach is particularly successful in that it pays proper respect to the two main ethical theories (deontology and utilitarianism) and combines them in a fruitful and promising way. However, their additional strategy of using empirical studies to mirror human moral decisions by considering as correct only those decisions that align with the majority view is misleading and seriously flawed. Rather, their system should be seen as a model of a descriptive study of ethical behaviour but not a model for normative ethics.

iii. Mixed Approaches: The Hybrid Approach

The hybrid model of human cognition (Wallach et al. 2010; Wallach and Allen 2010) combines a top-down component (theory-driven reasoning) and a bottom-up (shaped by evolution and learning) component that are considered the basis of both moral reasoning and decision-making. The result thus far is LIDA, an AGI software offering a comprehensive conceptual and computational model that models a large portion of human cognition. The hybrid model of moral reasoning attempts to re-create human decision-making by appealing to a complex combination of top-down and bottom-up approaches leading eventually to a descriptive but not a normative model of ethics. In addition, its somewhat idiosyncratic understanding of both approaches from moral philosophy does not in fact match how moral philosophers understand and use them in normative ethics. The model presented by Wallach et al. is not necessarily inaccurate with respect to how moral decision-making works in an empirical sense, but their approach is descriptive rather than normative in nature. Therefore, their empirical model does not solve the normative problem of how moral machines should act. Descriptive ethics and normative ethics are two different things. The former tells us how human beings make moral decisions; the latter is concerned with how we should act.

b. Autonomous Systems

The proposals for a system of machine ethics discussed in section 2.a. are increasingly being discussed in relation to autonomous systems the operation of which poses a risk of harm to human life. The two most-often discussed examples—which are at times discussed together and contrasted and compared with each other—are autonomous vehicles (also known as self-driving cars) and autonomous weapons systems (sometimes dubbed ‘killer robots’) (Purves et al. 2015; Danaher 2016; Nyholm 2018a).

Some authors think that autonomous weapons might be a good replacement for human soldiers (Müller and Simpson 2014). For example, Arkin (2009, 2010) argues that having machines fight our wars for us instead of human soldiers could lead to a decrease in war crimes if the machines were equipped with an ‘ethical governor’ system that would consistently follow the rules of war and engagement. However, others worry about the widespread availability of AI-driven autonomous weapons systems, because they think the availability of such systems might tempt people to go to war more often, or because they are sceptical about the possibility of an AI system that could interpret and apply the ethical and legal principles of war (see, for example, Royakkers and van Est 2015; Strawser 2010). There are also worries that ‘killer robots’ might be hacked (Klincewicz 2015).

Similarly, while acknowledging the possible benefits of self-driving cars—such as increased traffic safety, more efficient use of fuel and better-coordinated traffic—many authors have also

noted the possible accidents that could occur (Goodall 2014; Lin 2015; Gurney 2016; Nyholm 2018b, 2018c; Keeling 2020). The underlying idea is that autonomous vehicles should be equipped with ‘ethics settings’ that would help to determine how they should react to accident scenarios where people’s lives and safety are at stake (Gogoll and Müller 2017). This is considered another real-life application of machine ethics that society urgently needs to grapple with.

The concern for self-driving cars being involved in deadly accidents for which the AI system may not have been adequately prepared has already been realised, tragically, as some people have died in such accidents (Nyholm 2018b). The first instance of death while riding in an autonomous vehicle—a Tesla Model S car in ‘autopilot’ mode—occurred in May 2016. The first pedestrian was hit and killed by an experimental self-driving car, operated by the ride-hailing company Uber, in March 2018. In the latter case, part of the problem was that the AI system in the car had difficulty classifying the object that suddenly appeared in its path. It initially classified the victim as ‘unknown’, then as a ‘vehicle’, and finally as a ‘bicycle’. Just moments before the crash, the system decided to apply the brakes, but by then it was too late (Keeling 2020: 146). Whether the AI system in the car functions properly can thus be a matter of life and death.

Philosophers discussing such cases may propose that, even when it cannot brake in time, the car might swerve to one side (for example, Goodall 2014; Lin 2015). But what if five people were on the only side of the road the car could swerve onto? Or what if five people appeared on the road and one person was on the curb where the car might swerve? These scenarios are similar to the much-discussed ‘trolley problem’: the choice would involve killing one person to save five, and the question would become under what sorts of circumstances that decision would or would not be permissible. Several papers have discussed relevant similarities and differences between the ethics of crashes involving self-driving cars, on the one hand, and the philosophy of the trolley problem, on the other (Lin 2015; Nyholm and Smids 2016; Goodall 2016; Himmelreich 2018; Keeling 2020; Kamm 2020).

One question that has occupied ethicists discussing autonomous systems is what ethical principles should govern their decision-making process in situations that might involve harm to human beings. A related issue is whether it is ever acceptable for autonomous machines to kill or harm human beings, particularly if they do so in a manner governed by certain principles that have been programmed into or made part of the machines in another way. Here, a distinction is made between deaths caused by self-driving cars—which are generally considered a deeply regrettable but foreseeable side effect of their use—and killing by autonomous weapons systems, which some consider always morally unacceptable (Purves et al. 2015). Even a campaign has been launched to ‘stop killer robots’, backed by many AI ethicists such as Noel Sharkey and Peter Asaro.

One reason for arguing that autonomous weapons systems should be banned the campaign puts

forward is that what they call ‘meaningful human control’ must be retained. This concept is also discussed in relation to self-driving cars (Santoni de Sio and van den Hoven 2018). Many authors have worried about the risk of creating ‘responsibility gaps’, or cases in which it is unclear who should be held responsible for harm that has occurred due to the decisions made by an autonomous AI system (Matthias 2004; Sparrow 2007; Danaher 2016). The key challenge here is to come up with a way of understanding moral responsibility in the context of autonomous systems that would allow us to secure the benefits of such systems and at the same time appropriately attribute responsibility for any undesirable consequences. If a machine causes harm, the human beings involved in the machine’s action may try to evade responsibility; indeed, in some cases it might seem unfair to blame people for what a machine has done. Of course, if an autonomous system produces a good outcome, which some human beings, if any, claim to deserve praise for, the result might be equally unclear. In general, people may be more willing to take responsibility for good outcomes produced by autonomous systems than for bad ones. But in both situations, responsibility gaps can arise. Accordingly, philosophers need to formulate a theory of how to allocate responsibility for outcomes produced by functionally autonomous AI technologies, whether good or bad (Nyholm 2018a; Dignum 2019; Danaher 2019a; Tigard 2020a).

c. Machine Bias

Many people believe that the use of smart technologies would put an end to human bias because of the supposed ‘neutrality’ of machines. However, we have come to realise that machines may maintain and even substantiate human bias towards women, different ethnicities, the elderly, people with medical impairments, or other groups (Kraemer et al. 2011; Mittelstadt et al. 2016). As a consequence, one of the most urgent questions in the context of machine learning is how to avoid machine bias (Daniels et al. 2019). The idea of using AI systems to support human decision-making is, in general, an excellent objective in view of AI’s ‘increased efficiency, accuracy, scale and speed in making decisions and finding the best answers’ (World Economic Forum 2018: 6). However, machine bias can undermine this seemingly positive situation in various ways. Some striking cases of machine bias are as follows:

1. Gender bias in hiring (Dastin 2018);
2. Racial bias, in that certain racial groups are offered only particular types of jobs (Sweeney 2013);
3. Racial bias in decisions on the creditworthiness of loan applicants (Ludwig 2015);
4. Racial bias in decisions whether to release prisoners on parole (Angwin et al. 2016);
5. Racial bias in predicting criminal activities in urban areas (O’Neil 2016);
6. Sexual bias when identifying a person’s sexual orientation (Wang and Kosinski 2018);
7. Racial bias in facial recognition systems that prefer lighter skin colours (Buolamwini and Gebru 2018);
8. Racial and social bias in using the geographic location of a person’s residence as a proxy

for ethnicity or socio-economic status (Veale and Binns 2017).

We can recognise at least three reasons for machine bias: (1) data bias, (2) computational/algorithmic bias and (3) outcome bias (Springer et al. 2018: 451). First, a machine learning system that is trained using data that contain implicit or explicit imbalances reinforces the distortion in the data with respect to any future decision-making, thereby making the bias systematic. Second, a programme may suffer from algorithmic bias due to the developer's implicit or explicit biases. The design of a programme relies on the developer's understanding of the normative and non-normative values of other people, including the users and stakeholders affected by it (Dobbe et al. 2018). Third, outcome bias could be based on the use of historical records, for example, to predict criminal activities in certain particular urban areas; the system may allocate more police to a particular area, resulting in an increase in reported cases which would have been unnoticed before. This logic would substantiate the AI system's decision to allocate the police to this area, even though other urban areas may have similar or even greater numbers of crimes, more of which would go unreported due to the lack of policing (O'Neil 2016).

Most AI researchers, programmers and developers as well as scholars working in the field of technology believe that we will never be able to design a fully unbiased system. Therefore, the focus is on *reducing* machine bias and minimising its detrimental effects on human beings. Nevertheless, various questions remain. What type of bias cannot be filtered out and when should we be satisfied with the remaining bias? What does it mean for a person in court to be subject not only to human bias but also to machine bias, with both forms of injustice potentially helping to determine the person's sentence? Is one type of bias not enough? Should we not rather aim to eliminate human bias instead of introducing a new one?

d. The Problem of Opacity

AI systems are used to make many sorts of decisions that significantly impact people's lives. AI can be used to make decisions about who gets a loan, who is admitted to a university, who gets an advertised job, who is likely to reoffend, and so on. Since these decisions have major impacts on people, we must be able to understand the underlying reasons for them. In other words, AI and its decision-making need to be explainable. In fact, many authors discussing the ethics of AI propose explainability (also referred to as explicability) as a basic ethical criterion, among others, for the acceptability of AI decision-making (Floridi et al. 2018). However, many decisions made by an autonomous AI system are not readily explainable to people. This came to be called the problem of opacity.

The opacity of AI decision-making can be of different kinds, depending on relevant factors. Some AI decisions are opaque to those who are affected by them because the algorithms behind the decisions, though quite easy to understand, are protected trade secrets which the companies

using them do not want to share with anyone outside the company. Another reason for AI opacity is that most people lack the technical expertise to understand how an AI-based system works, even if there is nothing intrinsically opaque about the technology in question. With some forms of AI, not even the experts can understand the decision-making processes used. This has been dubbed the ‘black box’ problem (Wachter, Mittelstadt and Russell 2018).

On the individual level, it can seem to be an affront to a person’s dignity and autonomy when decisions about important aspects of their lives are made by machines if it is unclear—or perhaps even impossible to know—why machines made these decisions. On the societal level, the increasing prominence of algorithmic decision-making could become a threat to our democratic processes. Henry Kissinger, the former U.S. Secretary of State, once stated, ‘We may have created a dominating technology in search of a guiding philosophy’ (Kissinger 2018; quoted in Müller 2020). John Danaher, commenting on this idea, worries that people might be led to act in superstitious and irrational ways, like those in earlier times who believed that they could affect natural phenomena through rain dances or similar behaviour. Danaher has called this situation ‘the threat of algocracy’—that is, of rule by algorithms that we do not understand but have to obey (Danaher 2016b, 2019b).

But is AI opacity always, and necessarily, a problem? Is it equally problematic across all contexts? Should there be an absolute requirement that AI must in all cases be explainable? Scott Robbins (2019) has provided some interesting and noteworthy arguments in opposition to this idea. Robbins argues, among other things, that a hard requirement for explicability could prevent us from reaping all the possible benefits of AI. For example, he points out that if an AI system could reliably detect or predict some form of cancer in a way that we cannot explain or understand, the value of knowing the information would outweigh any concerns about not knowing how the AI system would have reached this conclusion. In general, it is also possible to distinguish between contexts where the procedure behind a decision matters in itself and those where only the quality of the outcome matters (Danaher and Robbins 2020).

Another promising response to the problem of opacity is to try to construct alternative modes of explaining AI decisions that would take into account their opacity but would nevertheless offer some form of explanation that people could act on. Sandra Wachter, Brent Mittelstadt, and Chris Russell (2019) have developed the idea of a ‘counterfactual explanation’ of such decisions, one designed to offer practical guidance for people wishing to respond rationally to AI decisions they do not understand. They state that ‘counterfactual explanations do not attempt to clarify how [AI] decisions are made internally. Instead, they provide insight into which external facts could be different in order to arrive at a desired outcome’ (Wachter et al. 2018: 880). Such an external, counterfactual way of explaining AI decisions might be a promising alternative in cases where AI decision-making is highly valuable but functions according to an internal logic that is opaque to most or all people.

e. Machine Consciousness

Some researchers think that when machines become more and more sophisticated and intelligent, they might at some point become spontaneously conscious as well (compare Russell 2019). This would be a sort of puzzling—but potentially highly significant from an ethical standpoint—side effect of the development of advanced AI. Some people are intentionally seeking to create machines with artificial consciousness. Kunihiro Asada, a successful engineer, set his goal as to create a robot that can experience pleasure and pain, on the basis that such a robot could engage in the kind of pre-linguistic learning that a human baby is capable of before it acquires language (Marchese 2020). Another example is Sophia the robot, whose developers at Hanson Robotics say that they wish to create a ‘super-intelligent benevolent being’ that will eventually become a ‘conscious, living machine’.

Others, such as Joanna Bryson, note that depending on how we define consciousness, some machines might already have some form of consciousness. Bryson argues that if we take consciousness to mean the presence of internal states and the ability to report on these states to other agents, then some machines might fulfil these criteria even now (Bryson 2012). In addition, Aïda Elamrani-Raoult and Roman Yampolskiy (2018) have identified as many as twenty-one different possible tests of machine consciousness.

Moreover, similar claims could be made about the issue of whether machines can have minds. If mind is defined, at least in part, in a functional way, as the internal processing of inputs from the external environment that generates seemingly intelligent responses to that environment, then machines could possess minds (Nyholm 2020: 145–46). Of course, even if machines can be said to have minds or consciousness in some sense, they would still not necessarily be anything like human minds. After all, the particular consciousness and subjectivity of any being will depend on what kinds of ‘hardware’ (such as brains, sense organs, and nervous systems) the being in question has (Nagel 1974).

Whether or not we think some AI machines are already conscious or that they could (either by accident or by design) become conscious, this issue is a key source of ethical controversy. Thomas Metzinger (2013), for example, argues that society should adopt, as a basic principle of AI ethics, a rule against creating machines that are capable of suffering. His argument is simple: suffering is bad, it is immoral to cause suffering, and therefore it would be immoral to create machines that suffer. Joanna Bryson contends similarly that although it is possible to create machines that would have a significant moral status, it is best to avoid doing so; in her view, we are morally obligated not to create machines to which we would have obligations (Bryson 2010, 2019). Again, this might all depend on what we understand by consciousness. Accordingly, Eric Schwitzgebel and Mara Garza (2015: 114–15) comment, ‘If society continues on the path towards developing more sophisticated artificial intelligence, developing a good theory of con-

sciousness is a moral imperative’.

Another interesting perspective is provided by Nicholas Agar (2019), who suggests that if there are arguments both in favour of and against the possibility that certain advanced machines have minds and consciousness, we should err on the side of caution and proceed on the assumption that machines do have minds. On this basis, we should then avoid any actions that might conceivably cause them to suffer. In contrast, John Danaher (2020) states that we can never be sure as to whether a machine has conscious experience, but that this uncertainty does not matter; if a machine behaves similarly to how conscious beings with moral status behave, this is sufficient moral reason, according to Danaher’s ‘ethical behaviourism’, to treat the machine with the same moral considerations with which we would treat a conscious being. The standard approach considers whether machines do actually have conscious minds and then how this answer should influence the question of whether to grant machines moral status (see, for example, Schwitzgebel and Garza 2015; Mosakas 2020; Nyholm 2020: 115–16).

f. The Moral Status of Artificial Intelligent Machines

Traditionally, the concept of moral status has been of utmost importance in ethics and moral philosophy because entities that have a moral status are considered part of the moral community and are entitled to moral protection. Not all members of a moral community have the same moral status, and therefore they differ with respect to their claims to moral protection. For example, dogs and cats are part of our moral community, but they do not enjoy the same moral status as a typical adult human being. If a being has a moral status, then it has certain moral (and legal) rights as well. The twentieth century saw a growth in the recognition of the rights of ethnic minorities, women, and the LGBTQ+ community, and even the rights of animals and the environment. This expanding moral circle may eventually grow further to include artificial intelligent machines once they exist (as advocated by the robot rights movement).

The notion of personhood (whatever that may mean) has become relevant in determining whether an entity has full moral status and whether, depending on its moral status, it should enjoy the full set of moral rights. One prominent definition of moral status has been provided by Frances Kamm (2007: 229):

So, we see that within the class of entities that count in their own right, there are those entities that *in their own right and for their own sake* could give us reason to act. I think that it is this that people have in mind when they ordinarily attribute moral status to an entity. So, henceforth, I shall distinguish between an entity’s counting morally in its own right and its having moral status. I shall say that *an entity has moral status when, in its own right and for its own sake, it can give us reason to do things such as not destroy it or help it.*

Things can be done for X's own sake, according to Kamm, if X is either conscious and/or able to feel pain. This definition usually includes human beings and most animals, whereas non-living parts of nature are mainly excluded on the basis of their lack of consciousness and inability to feel pain. However, there are good reasons why one should broaden their moral reasoning and decision-making to encompass the environment as well (Stone 1972, 2010; Atapattu 2015). For example, the Grand Canyon could be taken into moral account in human decision-making, given its unique form and great aesthetic value, even though it lacks personhood and therefore moral status. Furthermore, some experts have treated sentient animals such as great apes and elephants as persons even though they are not human (for example, Singer 1975; Cavalieri 2001; Francione 2009).

In addition, we can raise the important question of whether (a) current robots used in social situations or (b) artificial intelligent machines, once they are created, might have a moral status and be entitled to moral rights as well, comparable to the moral status and rights of human beings. The following three main approaches provide a brief overview of the discussion.

i. The Autonomy Approach

Kant and his followers place great emphasis on the notion of autonomy in the context of moral status and rights. A moral person is defined as a rational and autonomous being. Against this background, it has been suggested that one might be able to ascribe personhood to artificial intelligent machines once they have reached a certain level of autonomy in making moral decisions. Current machines are becoming increasingly autonomous, so it seems only a matter of time until they meet this moral threshold. A Kantian line of argument in support of granting moral status to machines based on autonomy could be framed as follows:

1. Rational agents have the capability to decide whether to act (or not act) in accordance with the demands of morality.
 - a. The ability to make decisions and to determine what is good *has* absolute value.
 - b. The ability to make such decisions *gives* rational persons absolute value.
2. A rational agent can act autonomously, including acting with respect to moral principles.
 - a. Rational agents have dignity *insofar* as they act autonomously.
 - b. Acting autonomously makes persons morally responsible.
3. Such a being—that is, a rational agent—has moral personhood.

It might be objected that machines—no matter how autonomous and rational—are not human beings and therefore should not be entitled to a moral status and the accompanying rights under a Kantian line of reasoning. But this objection is misleading, since Kant himself clearly states in his *Groundwork* (2009) that human beings should be considered as moral agents not because they are human beings, but because they are autonomous agents (Altman 2011; Timmermann

2020: 94). Kant has been criticised by his opponents for his logocentrism, even though this very claim has helped him avoid the more severe objection of speciesism—of holding that a particular species is morally superior simply because of the empirical features of the species itself (in the case of human beings, the particular DNA). This has been widely viewed as the equivalent of racism at the species level (Singer 2009).

ii. The Indirect Duties Approach

The indirect duties approach is based on Kant's analysis of our behaviour towards animals. In general, Kant argues in his *Lectures on Ethics* (1980: 239–41) that even though human beings do not have direct duties towards animals (because they are not persons), they still have indirect duties towards them. The underlying reason is that human beings may start to treat their fellow humans badly if they develop bad habits by mistreating and abusing animals as they see fit. In other words, abusing animals may have a detrimental, brutalising impact on human character.

Kate Darling (2016) has applied the Kantian line of reasoning to show that even current social robots should be entitled to moral and legal protection. She argues that one should protect life-like beings such as robots that interact with human beings when society cares deeply enough about them, even though they do not have a right to life. Darling offers two arguments why one should treat social robots in this way. Her first argument concerns people who witness cases of abuse and mistreatment of robots, pointing out that they might become 'traumatized' and 'de-sensitized'. Second, she contends that abusing robots may have a detrimental impact on the abuser's character, causing her to start treating fellow humans poorly as well.

Indeed, current social robots may be best protected by the indirect duties approach, but the idea that exactly the same arguments should also be applied to future robots of greater sophistication that either match or supersede human capabilities is somewhat troublesome. Usually, one would expect that these future robots—unlike Darling's social robots of today—will be not only moral patients but rather proper moral agents. In addition, the view that one should protect life-like beings 'when society cares deeply enough' (2016: 230) about them opens the door to social exclusion based purely on people's unwillingness to accept them as members of the moral community. Morally speaking, this is not acceptable. The next approach attempts to deal with this situation.

iii. The Relational Approach

Mark Coeckelbergh (2014) and David Gunkel (2012), the pioneers of the relational approach to moral status, believe that robots have a moral status based on their social relation with human beings. In other words, moral status or personhood emerges through social relations between different entities, such as human beings and robots, instead of depending on criteria inherent in

the being such as sentience and consciousness. The general idea behind this approach comes to the fore in the following key passage (Coeckelbergh 2014: 69–70):

We may wonder if robots will remain “machines” or if they can become companions. Will people start saying, as they tend to say of people who have “met their dog” ... , that someone has “met her robot”? Would such a person, having that kind of relation with that robot, still feel shame at all in front of the robot? And is there, at that point of personal engagement, still a need to talk about the “moral standing” of the robot? Is not moral quality already implied in the very relation that has emerged here? For example, if an elderly person is already very attached to her Paro robot and regards it as a pet or baby, then what needs to be discussed is that relation, rather than the “moral standing” of the robot.

The personal experience with the *Other*, that is, the robot, is the key component of this relational and phenomenological approach. The relational concept of personhood can be fleshed out in the following way:

1. A social model of autonomy, under which autonomy is not defined individually but stands in the context of social relations;
2. Personhood is absolute and inherent in every entity as a social being; it does not come in degrees;
3. An interactionist model of personhood, according to which personhood is relational by nature (but not necessarily reciprocal) and defined in non-cognitivist terms.

The above claims are not intended as steps in a conclusive argument; rather, they portray the general line of reasoning regarding the moral importance of social relations. The relational approach does not require the robot to be rational, intelligent or autonomous as an individual entity; instead, the social encounter with the robot is morally decisive. The moral standing of the robot is based on exactly this social encounter.

The problem with the relational approach is that the moral status of robots is thus based completely on human beings’ willingness to enter into social relations with a robot. In other words, if human beings (for whatever reasons) do not want to enter into such relations, they could deny robots a moral status to which the robots might be entitled on more objective criteria such as rationality and sentience. Thus, the relational approach does not actually provide a strong foundation for robot rights; rather, it supports a pragmatic perspective that would make it easier to welcome robots (who already have moral status) in the moral community (Gordon 2020c).

iv. The Upshot

The three approaches discussed in sections 2.f.i-iii. all attempt to show how one can make sense

of the idea of ascribing moral status and rights to robots. The most important observation is, however, that robots are entitled to moral status and rights independently of our opinion, once they have fulfilled the relevant criteria. Whether human beings will actually *recognise* their status and rights are a different matter.

g. Singularity and Value Alignment

Some of the theories of the potential moral status of artificial intelligent agents discussed in section 2.f. have struck some authors as belonging to science fiction. The same can be said about the next topic to be considered: singularity. The underlying argument regarding *technological singularity* was introduced by statistician I. J. Good in ‘Speculations Concerning the First Ultraintelligent Machine’ (1965):

Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion”, and the intelligence of man would be left far behind. Thus, the first ultraintelligent machine is the last invention that man need ever make.

The idea of an intelligence explosion involving self-replicating, super-intelligent AI machines seems inconceivable to many; some commentators dismiss such claims as a myth about the future development of AI (for example, Floridi 2016). However, prominent voices both inside and outside academia are taking this idea very seriously—in fact, so seriously that they fear the possible consequence of the so-called ‘existential risks’ such as the risk of human extinction. Among those voicing such fears are philosophers like Nick Bostrom and Toby Ord, but also prominent figures like Elon Musk and the late Stephen Hawking.

Authors discussing the idea of technological singularity differ in their views about what might lead to it. The famous futurist Ray Kurzweil is well-known for advocating the idea of singularity with exponentially increasing computing power, associated with ‘Moore’s law’, which points out that the computing power of transistors, at the time of writing, had been doubling every two years since the 1970s and could reasonably be expected to continue to do so in future (Kurzweil 2005). This approach sees the path to superintelligence as likely to proceed through a continuing improvement of the hardware. Another take on what might lead to superintelligence—favoured by the well-known AI researcher Stuart Russell—focuses instead on algorithms. From Russell’s (2019) point of view, what is needed for singularity to occur are conceptual breakthroughs in such areas as the studies of language and common-sense processing as well as learning processes.

Researchers concerned with singularity approach the issue of what to do to guard humanity against such existential risks in several different ways, depending in part on what they think these existential risks depend on. Bostrom, for example, understands superintelligence as consisting of a maximally powerful capacity to achieve whatever aims might be associated with artificial intelligent systems. In his much-discussed example (Bostrom 2014), a super-intelligent machine threatens the future of human life by becoming optimally efficient at maximising the number of paper clips in the world, a goal whose achievement might be facilitated by removing human beings so as to make more space for paper clips. From this point of view, it is crucial to equip super-intelligent AI machines with the right goals, so that when they pursue these goals in maximally efficient ways, there is no risk that they will extinguish the human race along the way. This is one way to think about how to create a beneficial super-intelligence.

Russell (2019) presents an alternative picture, formulating three rules for AI design, which might perhaps be viewed as an updated version of or suggested replacement for Asimov's fictional laws of robotics (see section 2.a.):

1. The machine's only objective is to maximise the realisation of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behaviour.

The theories discussed in this section represent different ideas about what is sometimes called 'value alignment'—that is, the concept that the goals and functioning of AI systems, especially super-intelligent future AI systems, should be properly aligned with human values. AI should be tracking human interests and values, and its functioning should benefit us and not lead to any existential risks, according to the ideal of value alignment. As noted in the beginning of this section, to some commentators, the idea that AI could become super-intelligent and pose existential threats is simply a myth that needs to be busted. But according to others, thinkers such as Toby Ord, AI is among the main reasons why humanity is in a critical period where its very future is at stake. According to such assessments, AI should be treated on a par with nuclear weapons and other potentially highly destructive technologies that put us all at great risk unless proper value alignment happens (Ord 2020).

A key problem concerning value alignment—especially if understood along the lines of Russell's three principles—is whose values or preferences AI should be aligned with. As Iason Gabriel (2020) notes, reasonable people may disagree on what values and interests are the right ones with which to align the functioning of AI (whether super-intelligent or not). Gabriel's suggestion for solving this problem is inspired by John Rawls' (1999, 2001) work on 'reasonable pluralism'. Rawls proposes that society should seek to identify 'fair principles' that could generate an overlapping consensus or widespread agreement despite the existence of more specific, reasonable disagreements about values among members of society. But how likely is it that this kind of con-

vergence in general principles would find widespread support? (See section 3.)

h. Other Debates

In addition to the topics highlighted above, other issues that have not received as much attention are beginning to be discussed within AI ethics. Five such issues are discussed briefly below.

i. AI as a form of Moral Enhancement or a Moral Advisor

AI systems tend to be used as ‘recommender systems’ in online shopping, online entertainment (for example, music and movie streaming), and other realms. Some ethicists have discussed the advantages and disadvantages of AI systems whose recommendations could help us to make better choices and ones more consistent with our basic values. Perhaps AI systems could even, at some point, help us improve our values. Works on these and related questions include Borenstein and Arkin (2016), Giubilini et al. (2015, 2018), Klincewicz (2016), and O’Neill et al. (2021).

ii. AI and the Future of Work

Much discussion about AI and the future of work concerns the vital issue of whether AI and other forms of automation will cause widespread ‘technological unemployment’ by eliminating large numbers of human jobs that would be taken over by automated machines (Danaher 2019a). This is often presented as a negative prospect, where the question is how and whether a world without work would offer people any prospects for fulfilling and meaningful activities, since certain goods achieved through work (other than income) are hard to achieve in other contexts (Gheaus and Herzog 2016). However, some authors have argued that work in the modern world exposes many people to various kinds of harm (Anderson 2017). Danaher (2019a) examines the important question of whether a world with less work might actually be preferable. Some argue that existential boredom would proliferate if human beings can no longer find a meaningful purpose in their work (or even their life) because machines have replaced them (Bloch 1954). In contrast, Jonas (1984) criticises Bloch, arguing that boredom will not be a substantial issue at all. Another related issue—perhaps more relevant in the short and medium-term—is how we can make increasingly technologised work remain meaningful (Smids et al. 2020).

iii. AI and the Future of Personal Relationships

Various AI-driven technologies affect the nature of friendships, romances and other interpersonal relationships and could impact them even more in future. Online ‘friendships’ arranged through social media have been investigated by philosophers who disagree as to whether rela-

tionships that are partly curated by AI algorithms, could be true friendships (Cocking et al. 2012; McFall 2012; Kiliarnta 2016; Elder 2017). Some philosophers have sharply criticised AI-driven dating apps, which they think might reinforce negative stereotypes and negative gender expectations (Frank and Klinecicz 2018). In more science-fiction-like philosophising, which might nevertheless become increasingly present in real life, there has also been discussion about whether human beings could have true friendships or romantic relationships with robots and other artificial agents equipped with advanced AI (Levy 2008; Sullins 2012; Elder 2017; Hauskeller 2017; Nyholm and Frank 2017; Danaher 2019c; Nyholm 2020).

iv. AI and the Concern About Human ‘Enfeeblement’

If more and more aspects of our lives are driven by the recommendations of AI systems (since we do not understand its functioning and we might question the propriety of its functioning), the results could include ‘a crisis in moral agency’ (Danaher 2019d), human ‘enfeeblement’ (Russell 2019), or ‘de-skilling’ in different areas of human life (Vallor 2015, 2016). This scenario becomes even more likely should technological singularity be attained, because at that point all work, including all research and engineering, could be done by intelligent machines. After some generations, human beings might indeed be completely dependent on machines in all areas of life and unable to turn the clock back. This situation is very dangerous; hence it is of utmost importance that human beings remain skilful and knowledgeable while developing AI capacities.

v. Anthropomorphism

The very idea of artificial intelligent machines that imitate human thinking and behaviour might incorporate, according to some, a form of anthropomorphising that ought to be avoided. In other words, attributing humanlike qualities to machines that are not human might pose a problem. A common worry about many forms of AI technologies (or about how they are presented to the general public) is that they are deceptive (for example, Boden et al. 2017). Many have objected that companies tend to exaggerate the extent to which their products are based on AI technology. For example, several prominent AI researchers and ethicists have criticised the makers of Sophia the robot for falsely presenting her as much more humanlike than she really is (for example, Sharkey 2018; Bryson 2010, 2019), and as being designed to prompt anthropomorphising responses in human beings that are somehow problematic or unfitting. The related question of whether anthropomorphising responses to AI technologies are always problematic requires further consideration, which it is increasingly receiving (for example, Coeckelbergh 2010; Darling 2016, 2017; Gunkel 2018; Danaher 2020; Nyholm 2020; Smids 2020).

This list of emerging topics within AI ethics is not exhaustive, as the field is very fertile, with new issues arising constantly. This is perhaps the fastest-growing field within the study of ethics and moral philosophy.

3. Ethical Guidelines for AI

As a result of widespread awareness of and interest in the ethical issues related to AI, several influential institutions (including governments, the European Union, large companies and other associations) have already tasked expert panels with drafting policy documents and ethical guidelines for AI. Such documents have proliferated to the point at which it is very difficult to keep track of all the latest AI ethical guidelines being released. Additionally, AI ethics is receiving substantial funding from various public and private sources, and multiple research centres for AI ethics have been established. These developments have mostly received positive responses, but there have also been some worries about the so-called ‘ethics washing’—that is, giving an ethical stamp of approval to something that might be, from a more critical point of view, ethically problematic (compare Tigard 2020b)—along with concerns that some efforts may be relatively toothless or too centred on the West, ignoring non-Western perspectives on AI ethics. This section, before discussing such criticisms, reviews examples of already published ethical guidelines and considers whether any consensus can emerge between these differing guidelines.

An excellent resource in this context is the overview by Jobin et al. (2019), who conducted a substantial comparative review of 84 sets of ethical guidelines issued by national or international organisations from various countries. Jobin et al. found strong convergence around five key principles—transparency, justice and fairness, non-maleficence, responsibility, and privacy, among many. Their findings are reported here to illustrate the extent of this convergence on some (but not all) of the principles discussed in the original paper. The number on the left indicates the number of ethical guideline documents, among the 84 examined, in which a particular principle was prominently featured. The codes Jobin et al. used are included so that readers can see the basis for their classification.

Ethical principle	Number of documents (N = 84)	Codes included
Transparency	73	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure
Justice and fairness	68	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access, distribution
Non-maleficence	60	Non-maleficence, security, safety, harm, protection, precaution, integrity (bodily or mental), non-subversion

Responsibility	60	Responsibility, accountability, liability, acting with integrity
Privacy	47	Privacy, personal or private information
Beneficence	41	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28	Trust
Sustainability	14	Sustainability, environment (nature), energy, resources (energy)
Dignity	13	Dignity
Solidarity	6	Solidarity, social security, cohesion

The review conducted by Jobin et al. (2019) reveals, at least with respect to the first five principles on the list, a significant degree of overlap in these attempts to create ethical guidelines for AI (see Gabriel 2020). On the other hand, the last six items on the list (beginning with beneficence) appeared as key principles in fewer than half of the documents studied. Relatedly, researchers working on the ‘moral machine’ research project, which examined people’s attitudes as to what self-driving cars should be programmed to do in various crash dilemma scenarios, also found great variation, including cross-cultural variation (Awad et al. 2018).

These ethical guidelines have received a fair amount of criticism—both in terms of their content and with respect to how they were created (for example, Metzinger 2019). For Metzinger, the very idea of ‘trustworthy AI’ is ‘nonsense’ since only human beings and not machines can be, or fail to be, trustworthy. Furthermore, the EU high-level expert group on AI had very few experts from the field of ethics but numerous industry representatives, who had an interest in toning down any ethical worries about the AI industry. In addition, the EU document ‘Ethical Guidelines for Trustworthy AI’ uses vague and non-confrontational language. It is, to use the term favoured by Resseguier and Rodrigues (2020), a mostly ‘toothless’ document. The EU ethical guidelines that industry representatives have supposedly made toothless illustrate the concerns raised about the possible ‘ethics washing’.

Another point of criticism regarding these kinds of ethical guidelines is that many of the expert panels drafting them are non-inclusive and fail to take non-Western (for example, African and Asian) perspectives on AI and ethics into account. Therefore, it would be important for future

versions of such guidelines—or new ethical guidelines—to include non-Western contributions. Notably, in academic journals that focus on the ethics of technology, there has been modest progress towards publishing more non-Western perspectives on AI ethics—for example, applying Dao (Wong 2012), Confucian virtue-ethics perspectives (Jing and Doorn 2020), and southern African relational and communitarian ethics perspectives including the ‘ubuntu’ philosophy of personhood and interpersonal relationships (see Wareham 2020).

4. Conclusion

The ethics of AI has become one of the liveliest topics in philosophy of technology. AI has the potential to redefine our traditional moral concepts, ethical approaches and moral theories. The advent of artificial intelligent machines that may either match or supersede human capabilities poses a big challenge to humanity’s traditional self-understanding as the only beings with the highest moral status in the world. Accordingly, the future of AI ethics is unpredictable but likely to offer considerable excitement and surprise.

5. References and Further Reading

- Agar, N. (2020). How to Treat Machines That Might Have Minds. *Philosophy & Technology*, 33(2): 269–82.
- Altman, M. C. (2011). *Kant and Applied Ethics: The Uses and Limits of Kant’s Practical Philosophy*. Malden, NJ: Wiley-Blackwell.
- Anderson, E. (2017). *Private Government: How Employers Rule Our Lives (and Why We Don’t Talk about It)*. Princeton, NJ: Princeton University Press.
- Anderson, M., and Anderson, S. (2011). *Machine Ethics*. Cambridge: Cambridge University Press.
- Anderson, S. L. (2011). Machine Metaethics. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics*, 21–27. Cambridge: Cambridge University Press.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias. In *ProPublica*, May 23. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL: CRC Press.
- Arkin, R. (2010). The Case for Ethical Autonomy in Unmanned Systems. *Journal of Military Ethics*, 9(4), 332–41.
- Asimov, I. (1942). *Runaround: A Short Story*. New York: Street and Smith.
- Asimov, I. (1986). *Robots and Empire: The Classic Robot Novel*. New York: HarperCollins.
- Atapattu, S. (2015). *Human Rights Approaches to Climate Change: Challenges and Opportunities*. New York: Routledge.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., and Rahwan, I. (2018). The Moral Machine Experiment. *Nature*, 563, 59–64.
- Bloch, E. (1985/1954). *Das Prinzip Hoffnung*, 3 vols. Frankfurt am Main: Suhrkamp.
- Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B., and Winfield, A. (2017). Principles of Robotics: Regulating Robots in the Real World. *Connection Science*, 29(2), 124–29.

- Borenstein, J. and Arkin, R. (2016). Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Science and Engineering Ethics*, 22, 31–46.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bryson, J. (2010). Robots Should Be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions*, 63–74. Amsterdam: John Benjamins.
- Bryson, J. (2012). A Role for Consciousness in Action Selection. *International Journal of Machine Consciousness*, 4(2), 471–82.
- Bryson, J. (2019). Patience Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics. *Ethics and Information Technology*, 20(1), 15–26.
- Buolamwini, J., and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*. PMLR, 81, 77–91.
- Čapek, K. (1920). *Rossum's Universal Robots*. Adelaide: The University of Adelaide.
- Cavalieri, P. (2001). *The Animal Question: Why Non-Human Animals Deserve Human Rights*. Oxford: Oxford University Press.
- Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York/Oxford: Oxford University Press.
- Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, 17, 7–65.
- Cocking, D., Van Den Hoven, J., and Timmermans, J. (2012). Introduction: One Thousand Friends. *Ethics and Information Technology*, 14, 179–84.
- Coeckelbergh, M. (2010). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology*, 12(3), 209–21.
- Coeckelbergh, M. (2014). The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology*, 27(1), 61–77.
- Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA and London: MIT Press.
- Copeland, B. J. (2020). Artificial Intelligence. *Britannica.com*. Retrieved from <https://www.britannica.com/technology/artificial-intelligence>.
- Danaher, J. (2016a). Robots, Law, and the Retribution Gap. *Ethics and Information Technology*, 18(4), 299–309.
- Danaher, J. (2016b). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, 29(3), 245–68.
- Danaher, J. (2019a). *Automation and Utopia*. Cambridge, MA: Harvard University Press.
- Danaher, J. (2019b). Escaping Skinner's Box: AI and the New Era of Techno-Superstition. Philosophical Disquisitions blog: <https://philosophicaldisquisitions.blogspot.com/2019/10/escaping-skinner-box-ai-and-new-era-of.html>.
- Danaher, J. (2019c). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*, 3(1), 5–24.
- Danaher, J. (2019d). The Rise of the Robots and the Crises of Moral Patience. *AI & Society*, 34(1), 129–36.
- Danaher, J. (2020). Welcoming Robots into the Moral Circle? A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, 26(4), 2023–49.
- Danaher, J., and Robbins, S. (2020). Should AI Be Explainable? Episode 77 of the Philosophical Disquisitions Podcast: <https://philosophicaldisquisitions.blogspot.com/2020/07/77-should-ai-be>

explainable.html.

- Daniels, J., Nkonde, M. and Mir, D. (2019). Advancing Racial Literacy in Tech. <https://datasociety.net/output/advancing-racial-literacy-in-tech/>.
- Darling, K. (2016). Extending Legal Protection to Social Robots: The Effects of Anthro- pomorphism, Empathy, and Violent Behavior towards Robotic Objects. In R. Calo, A. M. Froomkin and I. Kerr (eds.), *Robot Law*, 213–34. Cheltenham: Edward Elgar.
- Darling, K. (2017). “Who’s Johnny?” Anthropological Framing in Human–Robot Interaction, Integration, and Policy. In P. Lin, K. Abney and R. Jenkins (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, 173–92. Oxford: Oxford University Press.
- Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. Reuters, October 10. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
- Dehghani, M., Forbus, K., Tomai, E., and Klenk, M. (2011). An Integrated Reasoning Approach to Moral Decision Making. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics*, 422–41. Cambridge: Cambridge University Press.
- Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Berlin: Springer.
- Dobbe, R., Dean, S., Gilbert, T., and Kohli, N. (2018). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. In 2018 Workshop on Fairness, Accountability and Transparency in Machine Learning during ICMI, Stockholm, Sweden (July 18 version). <https://arxiv.org/abs/1807.00553>.
- Elamrani, A., and Yampolskiy, R. (2018). Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*, 26(5–6), 35–64.
- Elder, A. (2017). *Friendship, Robots, and Social Media: False Friends and Second Selves*. London: Routledge.
- Floridi, L. (2016). Should We Be Afraid of AI? Machines Seem to Be Getting Smarter and Smarter and Much Better at Human Jobs, yet True AI Is Utterly Implausible. Why? *Aeon*, May 9. <https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible>.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Francione, G. L. (2009). *Animals as Persons. Essay on the Abolition of Animal Exploitation*. New York: Columbia University Press.
- Frank, L., and Klinecicz, M. (2018): Swiping Left on the Quantified Relationship: Exploring the Potential Soft Impacts. *American Journal of Bioethics*, 18(2), 27–28.
- Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, available online at <https://link.springer.com/article/10.1007/s11023-020-09539-2>.
- Gheaus, A., and Herzog, L. (2016). Goods of Work (Other than Money!). *Journal of Social Philosophy*, 47(1), 70–89.
- Giubilini, A., and Savulescu, J. (2018). The Artificial Moral Advisor: The “Ideal Observer” Meets Artificial Intelligence. *Philosophy & Technology*, 1–20. <https://doi.org/10.1007/s13347-017-0285-z>.
- Gogoll, J., and Müller, J. F. (2017). Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*, 23(3), 681–700.

- Good, I. J. (1965). Speculations Concerning the First Ultrainelligent Machine. In F. Alt and M. Rubino (Eds.), *Advances in Computers*, vol. 6, 31–88. Cambridge, MA: Academic Press.
- Goodall, N. J. (2014). Ethical Decision Making during Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2424, 58–65.
- Goodall, N. J. (2016). Away from Trolley Problems and Toward Risk Management. *Applied Artificial Intelligence*, 30(8), 810–21.
- Gordon, J.-S. (2020a). Building Moral Machines: Ethical Pitfalls and Challenges. *Science and Engineering Ethics*, 26, 141–57.
- Gordon, J.-S. (2020b). What Do We Owe to Intelligent Robots? *AI & Society*, 35, 209–23.
- Gordon, J.-S. (2020c). Artificial Moral and Legal Personhood. *AI & Society*, online first at <https://link.springer.com/article/10.1007%2Fso0146-020-01063-2>.
- Guarini, M. (2006). Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems*, 21(4), 22–28.
- Gunkel, D. J., and Bryson, J. (2014). The Machine as Moral Agent and Patient. *Philosophy & Technology*, 27(1), 5–142.
- Gunkel, D. (2012). *The Machine Question. Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA: MIT Press.
- Gunkel, D. (2018). *Robot Rights*. Cambridge, MA: MIT Press.
- Gurney, J. K. (2016). Crashing into the Unknown: An Examination of Crash-Optimization Algorithms through the Two Lanes of Ethics and Law. *Alabama Law Review*, 79(1), 183–267.
- Himmelreich, J. (2018). Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice*, 21(3), 669–84.
- Himma, K., and Tavani, H. (2008). *The Handbook of Information and Computer Ethics*. Hoboken, NJ: Wiley.
- Jobin, A., Ienca, M., and Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Johnson, D. (1985/2009). *Computer Ethics*, 4th ed. New York: Pearson.
- Johnson, D., and Nissenbaum, H. (1995). *Computing, Ethics, and Social Values*. Englewood Cliffs, NJ: Prentice Hall.
- Jonas, H. (2003/1984). *Das Prinzip Verantwortung*. Frankfurt am Main: Suhrkamp.
- Kaliarnta, S. (2016). Using Aristotle’s Theory of Friendship to Classify Online Friendships: A Critical Counterpoint. *Ethics and Information Technology*, 18(2), 65–79.
- Kamm, F. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. Oxford, UK: Oxford University Press.
- Kamm, F. (2020). The Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and the Distribution of Harm. In S. M. Liao (Ed.) *The Ethics of Artificial Intelligence*, 79–108. New York: Oxford University Press.
- Kant, I. (1980). *Lectures on Ethics*, trans. Louis Infield, Indianapolis, IN: Hackett Publishing Company.
- Kant, I. (2009). *Groundwork of the Metaphysics of Morals*. New York: Harper Perennial Modern Classics.
- Keeling, G. (2020). The Ethics of Automated Vehicles. PhD Dissertation, University of Bristol. https://research-information.bris.ac.uk/files/243368588/Pure_Thesis.pdf.
- Kissinger, H. A. (2018). How the Enlightenment Ends: Philosophically, Intellectually—in Every Way—

Human Society Is Unprepared for the Rise of Artificial Intelligence. *The Atlantic*, June.

<https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/>.

- Klincewicz, M. (2016). Artificial Intelligence as a Means to Moral Enhancement. In *Studies in Logic, Grammar and Rhetoric*. <https://doi.org/10.1515/slgr-2016-0061>.
- Klincewicz, M. (2015). Autonomous Weapons Systems, the Frame Problem and Computer Security. *Journal of Military Ethics*, 14(2), 162–76.
- Kraemer, F., Van Overveld, K., and Peterson, M. (2011). Is There an Ethics of Algorithms? *Ethics and Information Technology*, 13, 251–60.
- Kurzweil, R. (2005). *The Singularity Is Near*. London: Penguin Books.
- Levy, D. (2008). *Love and Sex with Robots*. London: Harper Perennial.
- Lin, P. (2015). Why Ethics Matters for Autonomous Cars. In M. Maurer, J. C. Gerdes, B. Lenz and H. Winner (Eds.), *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, 69–85. Berlin: Springer.
- Lin, P., Abney, K. and Bekey, G. A. (Eds.). (2014). *Robot Ethics: The Ethical and Social Implications of Robotics. Intelligent Robotics and Autonomous Agents*. Cambridge, MA and London: MIT Press.
- Lin, P., Abney, K. and Jenkins, R. (Eds.) (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press.
- Loh, J. (2019). *Roboterethik. Eine Einführung*. Frankfurt am Main: Suhrkamp.
- Ludwig, S. (2015). Credit Scores in America Perpetuate Racial Injustice: Here's How. *The Guardian*, October 13. <https://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why>.
- Marchese, K. (2020). Japanese Scientists Develop “Blade Runner” Robot That Can Feel Pain. *Design Boom*, February 24. <https://www.designboom.com/technology/japanese-scientists-develop-hyper-realistic-robot-that-can-feel-pain-02-24-2020/>.
- Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 6(3), 175–83.
- McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. <http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf>.
- McFall, M. T. (2012). Real Character-Friends: Aristotelian Friendship, Living Together, And Technology. *Ethics and Information Technology*, 14, 221–30.
- Metzinger, T. (2013). Two Principles for Robot Ethics. In E. Hilgendorf and J.-P. Günther (Eds.), *Robotik und Gesetzgebung*, 263–302. Baden-Baden: Nomos.
- Metzinger, T. (2019). Ethics Washing Made in Europe. *Der Tagesspiegel*. <https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html>.
- Misselhorn, C. (2018). *Grundfragen der Maschinenethik*. Stuttgart: Reclam.
- Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. 3(2). <https://journals.sagepub.com/doi/full/10.1177/2053951716679679>.
- Mosakas, K. (2020). On the Moral Status of Social Robots: Considering the Consciousness Criterion. *AI & Society*, online first at <https://link.springer.com/article/10.1007/s00146-020-01002-1>.
- Müller, V. C., and Simpson, T. W. (2014). Autonomous Killer Robots Are Probably Good News. *Frontiers in Artificial Intelligence and Applications*, 273, 297–305.

- Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. *Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/entries/ethics-ai/>.
- Nyholm, S. (2018a). Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, 24(4), 1201–19.
- Nyholm, S. (2018b). The Ethics of Crashes with Self-Driving Cars: A Roadmap, I. *Philosophy Compass*, 13(7), e12507.
- Nyholm, S. (2018c). The Ethics of Crashes with Self-Driving Cars, A Roadmap, II. *Philosophy Compass*, 13(7), e12506.
- Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. London: Rowman and Littlefield.
- Nyholm, S., and Frank, L. (2017). From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? In J. Danaher and N. McArthur, *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.
- Nyholm, S., and Frank, L. (2019). It Loves Me, It Loves Me Not: Is It Morally Problematic to Design Sex Robots That Appear to Love Their Owners? *Techné: Research in Philosophy and Technology*, 23(3), 402–24.
- Nyholm, S., and Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem? *Ethical Theory and Moral Practice*, 19(5), 1275–89.
- Okyere-Manu, B. (Ed.) (2021). *African Values, Ethics, and Technology: Questions, Issues, and Approaches*. London: Palgrave MacMillan.
- O’Neil, C. (2016). *Weapons of Math Destruction*. London: Allen Lane.
- O’Neill, E., Klineciewicz, M. and Kemmer, M. (2021). Ethical Issues with Artificial Ethics Assistants. In C. Veliz (Ed.), *Oxford Handbook of Digital Ethics*. Oxford: Oxford University Press.
- Ord, T. (2020): *The Precipice: Existential Risk and the Future of Humanity*. London: Hachette Books.
- Picard, R. (1997). *Affective Computing*. Cambridge, MA and London: MIT Press.
- Purves, D., Jenkins, R. and Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice*, 18(4), 851–72.
- Rawls, J. (1999). *The Law of Peoples, with The Idea of Public Reason Revisited*. Cambridge, MA: Harvard University Press.
- Rawls, J. (2001). *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.
- Resseguier, A., and Rodrigues, R. (2020). AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics. *Big Data & Society*, online first at <https://journals.sagepub.com/doi/full/10.1177/2053951720942541>.
- Richardson, K. (2019). Special Issue: Ethics of AI and Robotics. *AI & Society*, 34(1).
- Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, 29(4), 495–514.
- Royakkers, L., and van Est, R. (2015). *Just Ordinary Robots: Automation from Love to War*. Boca Raton, FL: CRC Press.
- Russell, S. (2019). *Human Compatible*. New York: Viking Press.
- Ryan, M., and Stahl, B. (2020). Artificial Intelligence Guidelines for Developers and Users: Clarifying Their Content and Normative Implications. *Journal of Information, Communication and Ethics in Society*, online first at <https://www.emerald.com/insight/content/doi/10.1108/JICES-12-2019-0138/full/html>
- Santoni de Sio, F., and Van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A

- Philosophical Account. *Frontiers in Robotics and AI*. <https://www.frontiersin.org/articles/10.3389/frobt.2018.00015/full>.
- Savulescu, J., and Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In *Beyond Artificial Intelligence*, 79–95. Springer.
- Schwitzgebel, E., and Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119.
- Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioural and Brain Sciences*, 3(3), 417–57.
- Sharkey, Noel (2018), Mama Mia, It's Sophia: A Show Robot or Dangerous Platform to Mislead? *Forbes*, November 17. <https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#407e37877ac9>.
- Singer, P. (1975). *Animal liberation*. London, UK: Avon Books.
- Singer, P. (2009). Speciesism and Moral Status. *Metaphilosophy*, 40(3–4), 567–81.
- Smids, J. (2020). Danaher's Ethical Behaviourism: An Adequate Guide to Assessing the Moral Status of a Robot? *Science and Engineering Ethics*, 26(5), 2849–66.
- Smids, J., Nyholm, S. and Berkers, H. (2020). Robots in the Workplace: A Threat to—or Opportunity for—Meaningful Work? *Philosophy & Technology*, 33(3), 503–22.
- Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, 24(1), 62–77.
- Springer, A., Garcia-Gathright, J. and Cramer, H. (2018). Assessing and Addressing Algorithmic Bias – But Before We Get There. In *2018 AAAI Spring Symposium Series*, 450–54. <https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/viewPaper/17542>.
- Stone, C. D. (1972). Should Trees Have Standing? Toward Legal Rights for Natural Objects. *Southern California Law Review*, 45, 450–501.
- Stone, C. D. (2010). *Should Trees Have Standing? Law, Morality and the Environment*. Oxford: Oxford University Press.
- Strawser, B. J. (2010). Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles. *Journal of Military Ethics*, 9(4), 342–68.
- Sullins, J. (2012), Robots, Love, and Sex: The Ethics of Building a Love Machine. *IEEE Transactions on Affective Computing*, 3(4), 398–409.
- Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Acmqueue*, 11(3), 1–19.
- Tigard, D. (2020a). There is No Techno-Responsibility Gap. *Philosophy & Technology*, online first at <https://link.springer.com/article/10.1007/s13347-020-00414-7>.
- Tigard, D. (2020b). Responsible AI and Moral Responsibility: A Common Appreciation. *AI and Ethics*, online first at <https://link.springer.com/article/10.1007/s43681-020-00009-0>.
- Timmermann, J. (2020). *Kant's "Groundwork of the Metaphysics of Morals": A Commentary*. Cambridge: Cambridge University Press.
- Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–60.
- Vallor, S. (2015). Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology*, 28(1), 107–24.
- Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford University Press.
- Veale, M., and Binns, R. (2017). Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data. *Big Data & Society*, 4(2).

- Vinge, V. (1983). First Word. *Omni*, January, 10.
- Vinge, V. (1993). The Coming Technological Singularity. How to Survive in the Post-Human Era. *Whole Earth Review*, Winter.
- Wachter, S., Mittelstadt, B. and Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–87.
- Wallach, W., and Allen, C. (2010). *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press.
- Wallach, W., Franklin, S. and Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. *Topics in Cognitive Science*, 2(3), 454–85.
- Wang, Y., and Kosinski, M. (2018). Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images, *Journal of Personality and Social Psychology*, 114(2), 246–57.
- Wareham, C. S. (2020): Artificial Intelligence and African Conceptions of Personhood. *Ethics and Information Technology*, online first at <https://link.springer.com/article/10.1007/s10676-020-09541-3>
- Wong, P. H. (2012). Dao, Harmony, and Personhood: Towards a Confucian Ethics of Technology. *Philosophy & Technology*, 25(1), 67–86.
- World Economic Forum and Global Future Council on Human Rights (2018). How to Prevent Discriminatory Outcomes in Machine Learning (white paper). http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.

Author Information

John-Stewart Gordon
Email: johnstgordon@pm.me
Vytautas Magnus University
Lithuania

and

Sven Nyholm
Email: s.r.nyholm@uu.nl
The University of Utrecht
The Netherlands