Müller, Vincent C. (forthcoming 2019), 'Ethics of artificial intelligence and robotics', in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Palo Alto: CSLI, Stanford University). http://plato.stanford.edu/

Vincent C. Müller
TU Eindhoven & U Leeds & Alan Turing Institute
www.sophia.de

Vs. 0.96, 03.10.2019 (near-final draft, reviewed)

Ethics of Artificial Intelligence and Robotics

Artificial intelligence (AI) and robotics are technologies that seem to be of major importance for the development of humanity in the near future. They have raised fundamental questions about what we should do with these systems, what the systems themselves should do, and what risks they involve in the long term. They also challenge the human view of humanity as the sole intelligent and dominant species on Earth. The main division of this article is into issues that arise with AI systems as *objects*, i.e. tools made and used by humans (2), vs. AI systems as autonomous *subjects*, i.e. when ethics is for the AI systems themselves (3). The problem of a future 'singularity' or 'superintelligence' concerns both ethical use of AI and the ethics for AI systems – thus the separate final section (4).
The sections are ordered by ethical issue, rather than by technology: For each section we provide a general explanation of the *issue*, *positions* and *arguments*, then look how this plays out with current *technologies* and finally what *policy* consequences may be drawn. This means we have occasion to look at ethical issues that arise from utility of consequences, including 'risk', as well as issues that arise from a conflict with virtues, rules, or values. In sections (2.2) and (3.2) we also discuss questions of a more theoretical nature about 'artificial moral agents', in particular under which conditions an agent should be taken to have rights and responsibilities.

Ethics of Artificial Intelligence and Robotics

---

1 Introduction

1.1 Background of the Field

The ethics of AI and robotics is often focused on 'concerns' of various sorts – such concerns are characteristic of new technologies: Many of these concerns will turn out to be rather quaint (such as that trains are too fast for souls), some predictably wrong when they suggest that humans will change fundamentally (telephones will destroy personal communication, writing will destroy memory, or video cassettes will make going out redundant - https://xkcd.com/1289/), some predictably correct but moderately relevant (digital technology will destroy industries that make photographic film, cassette tapes, or LP records), but some broadly correct and relevant (such as that cars will kill children, change the landscape, and challenge good sleep). The task of an article such as this is to analyse the issues, and to deflate the non-issues - always keeping in mind that technologies are situated in a social and historical context. Some technologies, like nuclear power, cars or plastics, have caused ethical and political discussion and significant policy efforts to control the trajectory that these technologies are taking – usually once the damage is done. In addition to such 'ethical concerns', new technologies challenge current norms, conceptual systems and societal structures, which is of particular interest to philosophy. Finally, once we have understood a technology in its context, we need to shape our societal response, including regulation and law. All these concerns also exist in the case of the new technology of "artificial intelligence" or AI, and robotics – plus the more fundamental fear that it may end the era of human control on planet Earth.

In short, our philosophical task here is to present an orientation and analysis of the *issues*, the *positions* and the *arguments* in the ethics of AI and robotics – with an outlook on *policy*.

The ethics of AI and robotics have seen significant press coverage in recent years, which supports this kind of work, but also may end up undermining it: It often talks as if we already knew what would be ethical, and the issues are just what future technology will bring, and what we should do about it; as if the issues were only considerations of risk, security (Brundage et al. 2018), and the prediction of impact (e.g. in the job market). These are essentially technical problems, on how to achieve the outcome we know to be right; a problem of ethics, however, would require that we do not already know what is the right thing to do, perhaps because we do not know which are the factors that matter or because there is a tension between different values or consequences. In this sense, deception, theft and killing with AI are not a problem for ethics, but whether these are permissible under certain circumstances is a problem. Therefore, this article does not follow much of the current discussion in policy and industry with its focus on image and public relations – where the label "ethical" is really not much more than the new "green". Instead it focuses on the genuine problems of ethics where we do not readily know what the answers are.

A caveat is in order for our presentation: The ethics of AI and robotics is a very young field within applied ethics, with significant dynamics but few well-established issues and no authoritative overviews, though there is a promising outline (European Group on Ethics in Science and New Technologies 2018), and there are beginnings on societal impact (Floridi et al. 2018; Taddeo and Floridi 2018; Taylor et al. 2018; Walsh 2018; Bryson 2019; Gibert 2019; SIENNA 2019; Whittlestone et al. 2019), sometimes with specific policy recommendations (AI HLEG 2019; IEEE 2019). So this article cannot just reproduce what the community has achieved thus far, but must propose an ordering where little order exists, including the identification of promising lines for future work – but without spending much time on the historical development of the field.

1.2 AI & Robotics

The notion of 'artificial intelligence' (AI) is understood broadly here, as any kind of artificial computational system that shows intelligent behaviour, i.e. complex behaviour that is conducive to reaching goals. In particular, we do not wish to restrict 'intelligence' to what would require intelligence if done by *humans*, as Minsky said (1985) and is often repeated (Taddeo and Floridi 2018, 751). This means we can incorporate machines from 'technical AI' that show only limited abilities in learning or reasoning but excel at the automation of particular tasks, as well as 'general AI' that aims at creating a generally intelligent agent. However, AI somehow goes deeper than other technologies since it is the project of AI to create machines that have a feature central to how we humans see ourselves, namely as feeling, thinking, intelligent beings – thus the field of 'philosophy of AI'. The main purposes of an artificial intelligent agent probably involve sensing, modelling, planning and action, but current AI applications also include perception and sensing, text analysis, natural language processing (NLP), logical reasoning, game-playing, decision support systems, data analytics, predictive analytics, as well as autonomous vehicles and other forms of robotics (P. Stone et al. 2016). AI may involve any number of computational techniques to achieve these aims, be that classical symbol-manipulating AI, be it inspired by natural cognition, or machine learning via neural networks – the area that currently looks most dynamic (Goodfellow, Bengio, and Courville 2016; Silver et al. 2018). Some of the ethical issues we discuss here apply to the whole field of AI, and some only to sub-fields.

Historically, it is remarkable that the term "AI" used to be very broad approximately 1950-1980, then it came into disrepute during the 'AI winter', ca. 1975-1995, and narrowed: 'machine learning', 'natural language processing' and any 'data science' were often not labelled as 'AI'. It is now since, ca. 2010, that the use broadened again, and at times almost all of computer science and cognitive science is lumped under 'AI' – now a name to be proud of, on the edge of hype again, and a booming industry with massive capital investment (Shoham et al. 2018) that, as politicians say, "… promises to drive growth of the ... economy, enhance our economic and national security, and improve our quality of life." (Trump 2019, 1).

While AI can be entirely software and data processing, robots are physical machines with actuators that move and interact with the environment, that exert physical force on the world, such as a gripper or a turning wheel. From this point of view, autonomous cars or planes are robots, and only a minuscule portion of robots is human-shaped or 'humanoid', like in the movies. Some robots use AI, and some do not – e.g. typical industrial robots (of which there are millions) blindly follow completely defined scripts with minimal sensory input and no learning or reasoning that would add complexity. It is probably fair to say that while robotics systems cause more concerns in the general public, AI systems are actually more likely to have a greater impact on humanity. Also, systems with for a narrow set of tasks are less likely to cause new issues than systems that are more flexible and autonomous.

The fields of robotics and AI can thus be seen as two overlapping circles of systems: systems that are only AI, systems that are only robotics, and systems that are both. The scope of this article is not just the intersection, but the union of both circles.

1.3 A Note on Policy

There is significant public discussion about AI ethics and frequent pronouncements from politicians that the matter requires new policy – however, this is easier said than done: Actual technology policy is difficult to plan and to enforce. Technology policy can take many forms, from incentives and funding, infrastructure, taxation, good-will statements, regulation by various actors (self-regulation, local, national, international), to law. Policy for AI will possibly come into conflict with other aims of technology policy (e.g. sustainable development) or general policy (e.g. economic growth). One important practical aspect is which agents are involved in the development of a policy and what the power structures are. For people who work in in ethics and policy, there is probably a tendency to overestimate the impact and threats from any new technology, and to underestimate how far current regulation can reach (e.g. for product liability). Governments, parliaments, associations and industry circles in Europe and North-America have produced reports and white papers in recent years (we maintain a list on PT-AI Policy Documents and Institutions), some have generated good-will slogans ('trusted/responsible/humane/human-centred/good/beneficial AI'), but, as of early 2019, very little actual policy has been produced – beyond funding for AI, and for AI policy.

There are beginnings: The latest EU policy document suggests 'trustworthy AI' should be lawful, ethical and technically robust, and then spells this out as seven requirements: human oversight, technical robustness, privacy and data governance, transparency, fairness, well-being and accountability (AI HLEG 2019). Much European research now runs under the slogan 'responsible research and innovation' (RRI) and 'technology assessment' is a standard field since the advent of nuclear power. Professional ethics is now a standard field in Information Technology as well, and this includes issues relevant here (e.g. confidentiality). There is a risk that the current vision statements and self-regulation in the industry tend to delegate the decisions to experts, "a narrow circle of who can or should adjudicate ethical concerns around AI/ML" (Greene, Hoffmann, and Stark forthcoming), rather than incorporating societal stakeholders more deeply. A useful summary of an ethical framework for AI is given in (European Group on Ethics in Science and New Technologies 2018, 13ff). On general AI policy, see (Calo 2018) as well as (Crawford and Calo 2016; Stahl, Timmermans, and Mittelstadt 2016; Johnson and Verdicchio 2017; Giubilini and Savulescu 2018).

On wider societal implications (Jacobs et al. 2019) have useful perspectives. The more political angle of technology is often discussed in 'science and technology studies' (STS), and sadly rather separated from philosophy or ethics. As books like *The Ethics of Invention* (Jasanoff 2016) show, the concerns are often quite similar. In this article, we discuss the theoretical approaches and the policy for each type of issue separately, rather than for AI or robotics in general.

2 Ethics for the Use of AI & Robotics Systems

In this section we outline the ethical issues of human use of AI and robotics systems that can be more or less autonomous – which means we look at issues that arise with certain uses, and would not arise with others, though it must be kept in mind that technologies will always cause some uses to be easier and thus more frequent, and hinder other uses: the technology is not ethically neutral. The design of technical artefacts has ethical relevance for their use as well (Houkes and Vermaas 2010; Verbeek 2011), so beyond 'responsible use', we also need 'responsible research and innovation' in this field.

There are authors, however, who prefer to see the ethics of AI and robotics generally from the perspective of machine ethics, rather than from the perspective of use – e.g. (Dignum 2018) suggests the field is about three things: ethics by design, ethics in design, and ethics for design. This also stresses the need to look at different stages of the systems' life cycle.

We will discuss both what is ethically permitted (or not) and what is ethically required (or not). The focus on use does not pre-judge what kinds of approaches are best suited towards tackling these issues; they might well be virtue ethics (Vallor 2017) rather than consequentialist or value-based (Floridi et al. 2018). This section is neutral with respect to the question whether AI systems truly have 'intelligence' or other mental properties: It would apply equally well if AI and robotics are merely seen as the current face of automation. Useful surveys for the ethics of robotics include (Calo, Froomkin, and Kerr 2016; Royakkers and van Est 2016; Lin, Abney, and Jenkins 2017) on robot law and (Tzafestas 2016), with a chapter on the cultural background of robot ethics in Japan.

2.1 Privacy, Surveillance & Manipulation

2.1.1 Privacy & Surveillance

There is a general discussion about privacy and surveillance, which is largely independent of AI technology (e.g. Macnish 2017; Roessler 2017), and mainly concerns the access to private data and data that is personally identifiable. Privacy has several well recognised aspects, e.g. 'the right to be let alone', information privacy, privacy as an aspect of personhood, control over information about me, right to secrecy (Bennett and Raab 2006). Privacy studies have historically focused on state surveillance by secret services, which has significantly increased in the digital age (some of US state surveillance became more widely known with the 'Snowden revelations' in 2013). State surveillance still continues with many states comprehensively spying on communication and location metadata of their own citizens, and on communication and location data of citizens of other states – and then sometimes exchanging that information with the state of that citizen (e.g. in the 'five eyes' network: Australia, Canada, New Zealand, the UK and the USA). Generally data collection and storage is now all digital, our lives are more and more digital, most digital data is connected to a single Internet, and there is more and more sensor technology around that generates data (sound, video, movement, temperature, …). Every new AI technology amplifies the known issues, e.g. face recognition in photos and surveillance material allows identification and thus profiling and searching for individuals (Whittaker et al. 2018, 15ff). AI massively increases the possibilities of intelligent data collection and surveillance, and the possibilities for data analysis and for finding patterns. This increase applies to state agents, but increasingly to private businesses as well – some corporations, such as Google,

Microsoft or Facebook, have a fairly complete access to individual and network data. Businesses and individuals also trade data from various sources with anyone who will pay, including the police. At the same time, the control over who collects which data, and who has access, is much harder in the digital world than it was in the analogue world of letters and telephone calls.
Robotic devices have not yet played a major role in this area, except for security patrolling, but this will change once they are more common outside industry. Robotic devices together with the 'Internet of things', the so-called 'smart' systems (phone, TV, oven, lamp, home, …), the 'smart city' (Sennett 2018) and 'smart governance', are set to become part of the data-gathering machinery that offers more detailed data, of different types, in real time, with ever more information – to whomever has access.
It is clear that these systems will often reveal facts about humans that we themselves wish to suppress or are not aware of: They know more about us than we know ourselves. Even just observing online behaviour allows insights into our mental states (Burr and Christianini forthcoming) and manipulation (see below (2.1.2)), which has led to calls for protection of inferences or 'derived data' (Wachter and Mittelstadt forthcoming). With the last sentence of his bestselling book (Harari 2016) asks about what he thinks of as long-term consequences of AI: "What will happen to society, politics and daily life when non-conscious but highly intelligent algorithms know us better than we know ourselves?"
Privacy-preserving techniques who can conceal the identity of persons or groups to a large extent are now a standard staple in data science; they include (relative) anonymisation, access control (plus encryption), differential privacy and other models where computation is carried out without access to full non-encrypted input data (Stahl and Wright 2018). With more effort and cost, such techniques can avoid many of the privacy issues, though probably not all, and once the data has been collected and stored, the problem remains. Some companies, notably Apple, have also seen better privacy as a competitive advantage that can be leveraged and sold at a price to their consumers.
Data analysis is often used in 'predictive analytics' in business, healthcare and other fields, to foresee future developments. One such use is in 'predictive policing' (Programs 2014), which many fear might lead to an erosion of public liberties (Ferguson 2017) because 'predictive analytics' can give power to people using it, and take away power from the predicted. It appears, however, that the worries about policing lives to some extent off futuristic scenarios where law enforcement foresees and punishes planned actions, rather than waiting until a crime has been committed (like in the 2002 film 'Minority Report'). One concern is that these systems might perpetuate bias that was already in the data used to set up the system, e.g. by increasing the level of control of a particular population and then finding more crime in that population (for such issues, see 2.2 below). Actual 'predictive policing' or 'intelligence led policing' techniques mainly concern the question of how police forces are most efficiently applied by predicting where they will be needed, who or what should be controlled – which is something an experienced police force will have done to some extent. Also, police officers can be provided with more data that allows more control and better decisions in workflow support software (e.g. 'ArcGIS'). Whether this is problematic depends on the appropriate level of trust in the technical quality of these systems, and on the evaluation of aims of the police work. Perhaps a recent paper title points in the right direction: "AI ethics in predictive policing: From models of threat to an ethics of care" (Asaro 2019).
Legally, one of the major difficulties is to actually enforce regulation, both on the level of the state and on the level of the individual who has a claim – but must identify the responsible legal entity, prove the action, perhaps prove intent, find a court that declares itself competent … and eventually get the court to actually enforce its decision. Well-established legal protection of rights such as consumer rights, product liability and other civil liability or protection of intellectual property rights is often missing in digital products, or hard to enforce. This means that companies with a 'digital' background are used to testing their products on the consumers, without fear of liability, while heavily defending their intellectual property rights. This 'Internet Libertarianism' is sometimes taken to assume that technical solutions will take care of societal problems by themselves (Mozorov 2013).
The policy and regulation in these areas in favour of civil liberties and protection of individuals is under very intense pressure from businesses lobbying, from secret services and other state agencies that live off surveillance (see the next section). As a result, privacy protection has diminished massively as compared to the pre-digital age where communication was based on letters, analogue telephone communications, and personal conversation – and surveillance involved individual people 'listening in', usually under significant legal constraints. It is probably fair to say that as individuals and groups we have lost control of our data, and thus perhaps of ourselves (see the next section).

2.1.2 Manipulation of Behaviour

Manipulation of online behaviour is probably a core way of doing business on the Internet at the moment. The data trail we leave behind is how our 'free' services are paid for, enabled by AI for the analysis or 'valuation' of that data – but we are not told about that data collection and its value, and we are manipulated into leaving ever more such data. The primary focus of social media, gaming, and most of the Internet in this 'surveillance economy' is to gain, maintain and direct attention – and thus data supply. Essentially, this is a surveillance and attention economy: "Surveillance is the business model of the Internet" (Schneier 2015) and its economic system is "surveillance capitalism" (Zuboff 2019). The result is that "In this vast ocean of data, there is a frighteningly complete picture of us" (Smolan 2016, 1:01). This business model of corporations like Google or Facebook appears to be based on exploiting human weaknesses, deception, furthering procrastination, generating addiction, and manipulation (Harris 2016) – despite several claims to the contrary. Surveillance capitalism has caused many attempts to escape from the grasp of the 'big 5' (Amazon, Google/Alphabet, Microsoft, Apple, Facebook) or 'social media', e.g. in exercises of digital chastity or 'minimalism' (Newport 2019), or through the open source movement, but it appears that present-day computer users have lost their autonomy to escape from the 'big 5' while continuing with their life and work. We may also have lost ownership of our data, if 'ownership' is the right relation here.

The issues of AI in surveillance go beyond the mere *accumulation* of data and direction of attention: They include the *use* of information to manipulate behaviour, online and offline, in a way that undermines autonomous rational choice. Of course efforts to manipulate behaviour are ancient, but it may be that these gain a new quality in AI systems based on big-data analytics.

Given the intense interaction with data systems and the deep knowledge about individuals, the users are vulnerable to 'nudges', manipulation and deception. Given sufficient prior data, algorithms can be used to provide individuals or small groups with the kind of targeted input that is likely to influence these particular individuals.

Predictably, advertisers, marketers and online sellers will use any legal means at their disposal to make us part with more of our money, including exploitation of behavioural biases and deception (Costa and Halpern 2019) – e.g. through 'dark patterns' on web pages or in games (Mathur et al. 2019). Such manipulation is the business model in much of the gambling and gaming industries, but it is spreading, e.g. to low-cost airlines and, ultimately, maximum profit is the aim (or a central aim) of businesses like the 'big 5'.

Furthermore, social media are now the prime locations for political propaganda, both in more democratic systems and in less democratic ones. This influence can be used to influence voting behaviour, as in the Facebook-Cambridge Analytica 'scandal' (Woolley and Howard 2017; Bradshaw, Neudert, and Howard 2019) and it may harm the autonomy of individuals (Susser, Roessler, and Nissenbaum 2019).

Furthermore, improved AI faking technologies make what once was reliable evidence into unreliable evidence – this has already happened to digital photos, sound recordings and video … and it will soon be quite easy to create (rather than alter) 'deep fake' text, photos and video material with any content. OpenAI showed a similarity text creator GPT-2 early in 2019 but refrained from opening its code. (A simple version of this phenomenon are machine-driven Twitter accounts, some imitating actually existing people.) Soon, sophisticated real-time interaction with persons over texting, phone or video will be faked, too. So we cannot trust any digital interaction, while we are at the same time increasingly dependent on such interaction.

One more specific issue is that so-called 'machine learning' techniques in AI rely on training with vast amounts of data, so their performance requires that such data is available. This means there will often be a trade-off between privacy and rights to data vs. technical quality of the product. This influences the consequentialist evaluation of privacy-violating practices: Better data for higher quality training is available if privacy is reduced – though there may be other ways to achieve high-quality data sets, with more effort.

The policy in this field has its ups and downs; while the EU General Data Protection Regulation (GDPR) (EU Parliament 2016) has strengthened privacy protection, the US and China prefer growth with less regulation (N. Thompson and Bremmer 2018), likely in the hope that this gives them a competitive advantage in machine learning technologies that require very large data sets. It is clear that state, business and private actors have increased their ability to invade privacy with the help of AI technology and will continue to do so to further their particular interests – unless reined in by policy in the interest of society. (The UK has launched a "Centre for Data Ethics and Innovation Consultation" late in 2018.) Public control of state and private actors in this area is still minimal.

2.2 Our Epistemic Condition: Opacity and Bias

Automated AI decision support systems and 'predictive analytics' operate on data and produce a decision as 'output'. This output may be relatively trivial like "this restaurant matches your preferences", "the patient in this X-ray has completed bone growth", or have greater significance for a person, e.g. if it says "application to credit card declined", "donor organ will be given to another patient", "bail is denied", or "target identified and engaged". These systems raise "significant concerns about lack of due process, accountability, community engagement, and auditing" (Whittaker et al. 2018, 18ff). They are are part of a power structure where "we are creating decision-making processes that constrain and limit opportunities for human participation" (Danaher 2016b, 245).

At the same time, it will often be impossible for the affected person to know how the system came to this output, i.e. the system is 'opaque' to the person. If the system involves machine learning, it will typically be opaque even to the expert how a particular pattern was identified, or even what the pattern is. Furthermore, the decision system might have a bias and produce unfair decisions. Bias in decision systems and data sets is exacerbated by these sets and systems not being epistemically readily accessible, or 'opaque'. So, at least in the cases where there is a desire to remove bias, the analysis of opacity and bias has to go hand in hand, and the political response has to tackle both issues together.

2.2.1 Opacity of AI Systems

Many AI systems rely on machine learning techniques in (simulated) neural networks that will extract patterns from a given dataset, with or without 'correct' solutions provided, i.e. supervised, semi-supervised or unsupervised. With these techniques, the 'learning' captures patterns in the data and these are labelled in a way that appears useful to the programmer (e.g. "male" or "insecure") while the programmer does not really know which patterns in the data the system has used. In fact the programs are evolving, so when new data comes in, or new feedback is given ("this was correct", "this was incorrect"), the patterns used by the learning system change – depending on the type of machine learning system. What this means is that the outcome cannot really be explained, it is not transparent to the user or programmers, it is opaque. Furthermore, the quality of the program depends heavily on the quality of the data provided, following the old slogan "garbage in, garbage out". So, if the data already involved a bias (e.g. police data about the skin colour of suspects, or job data including gender), then the program will reproduce that bias. There are proposals for a standard description of datasets in a 'datasheet' that would make the identification of such bias more feasible (Gebru et al. 2018). There is a significant recent literature about the limitations of machine learning systems (Marcus 2018), that are essentially sophisticated data filters. This technology is likely at the peak of the 'hype cycle' at the moment. Opacity is a central issue in what is now often called 'data ethics' (Floridi and Taddeo 2016; Mittelstadt and Floridi 2016), particularly in the ethics of big data – whether AI is used there, or not.

In the EU, some of these issues have been taken into account with the GDPR, which foresees that consumers who are faced with a decision based on data processing (e.g. on a loan application) have a legal "right to explanation" – how far this goes and to what extent it can be enforced is disputed, however (Goodman and Flaxman 2016; Wachter, Mittelstadt, and Floridi 2017; Wachter, Mittelstadt, and Russell 2018). How far 'explicability' should go is an open question, given that the abilities of humans to explain and provide reasons are not too impressive - (Zerilli et al. forthcoming) argue there may be a double standard here where we demand too much of machine-based decisions.

There are several technical activities that aim at 'explainable AI', starting with (Van Lent, Fisher, and Mancuso 1999; Lomas et al. 2012) and, more recently, a DARPA programme (Gunning 2018) and the AI4EU project on 'human-centred AI'. More broadly, the demand for "a mechanism for elucidating and articulating the power structures, biases, and influences that computational artefacts exercise in society" (Diakopoulos 2015) is sometimes called "algorithmic accountability reporting". This does not mean that we expect an AI to 'explain its reasoning' – doing so would require far more serious moral autonomy than we currently attribute to AI systems (see below 3.2).

Kissinger pointed out that there is a fundamental problem for democratic decision-making if we rely on a system that is supposedly superior to mere humans, but cannot explain its decisions. He says we may have "generated a potentially dominating technology in search of a guiding philosophy" (Kissinger 2018). (Danaher 2016b) calls this problem the 'algocracy'. In a similar vein, (Cave 2019) stresses that we need a broader societal move towards more 'democratic' decision-making to avoid AI being a force that leads to a Kafka-style impenetrable suppression system in public administration and elsewhere. This general political angle of this discussion has been stressed by (O'Neil 2016) in her influential book *Weapons of Math Destruction,* and in (Yeung and Lodge 2019).

2.2.2 Bias in Decision Systems

Bias typically surfaces when unfair judgments are made because the individual making the judgment is influenced by a characteristic that is actually irrelevant to matter at hand, typically a discriminatory preconception about members of a group; e.g. they have a tendency underestimate a person's intellectual ability due to information about gender or skin colour. So the first form of bias is a cognitive feature of a person, learned in a particular social context at a particular time. It is often not make explicit and at times the person concerned is not conscious of having that bias – they may even be honestly and explicitly opposed to a bias they are found to have (e.g. through priming, cf. (Graham and Lowery 2004)). In some jurisdictions there are 'protected characteristics' that may not be used for distinction at all, i.e. their use is always considered discrimination (unless there is a specified exception in the law) – in the (UK Equality Act 2010) these would be: age, disability, gender reassignment, marriage or civil partnership, pregnancy or maternity, race (including colour, nationality, ethnic or national origin), religion or belief, sex, sexual orientation. This can lead to problems, e.g. when an insurance company finds out that these characteristics are predictive of higher risk for causing car accidents. On fairness vs. bias in machine learning, see (Binns 2018).

Apart from the social phenomenon of learned bias, the human cognitive system is generally prone to have various kinds of 'cognitive biases' (a 2ND form), e.g. the 'confirmation bias' where humans tend to interpret information as confirming what they already believe. Cognitive biases are said to impede performance in rational judgment (Kahnemann 2011) – though at least some cognitive biases generate an evolutionary advantage, e.g. economical use of resources for intuitive judgment. There is a question whether AI systems could or should have such cognitive bias.

A third form is bias in present in data, when it exhibits systematic error, e.g. one of the various kinds of 'statistical bias'. Sometimes, it can be detected automatically in human language corpora (Caliskan, Bryson, and Narayanan 2017). Strictly, any given dataset will only be unbiased for a particular kind of issue, so the mere creation of a dataset involves the danger that it be used for a different kind of issue, and then turn out biased for that kind. AI systems are often used to create such data, e.g. in the selection of news displayed to a particular user, creating a 'filter bubble'. Machine learning uses datasets for training and testing that are often parts of the same larger set, meaning they might carry the same bias – we might call this 'historical bias'. Bias can thus be present but unknown, which makes it harder to fight. Automation on the basis of such data would then just not fix the bias, but codify and automate the bias.

The fourth type of bias would thus be an AI decision system that has a cognitive feature similar to a biased human (be this due to learning from biased data or from other sources); this is sometimes called 'algorithmic bias'. This is difficult to avoid in backwards-looking machine learning systems because they can only learn from the past. Such historical bias was discovered in an automated recruitment screening system at Amazon (discontinued early 2017) that discriminated against women – presumably because the company had a history of hiring women less frequently than they occurred in the applicant pool. Historical bias can also be an issue in 'predictive policing', see section (2.1.1). For example, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a system to predict whether a defendant would re-offend, was found to be as successful (65.2% accuracy) as a group of random humans, or a simple linear predictor (Dressel and Farid 2018). Both COMPAS and the human group produced more false positives and less false negatives for black defendants. The problem with such systems is thus bias and excessive/insufficient trust. Such bias can then be reinforced, e.g. if black suspects are controlled by police more often for some offence, and are thus more likely to be found actually guilty of that offence. The political dimensions of such automated systems in the USA are investigated in (Eubanks 2018).

There are significant technical efforts made to detect and remove bias from AI systems, but it is fair to say that these are in early stages: see UK Institute for Ethical AI & Machine Learning, (Brownsword, Scotford, and Yeung 2017; Yeung and Lodge 2019). It appears that technological fixes have their limits in that they need a mathematical notion of fairness, which is hard to come by (Whittaker et al. 2018, 24ff; Selbst et al. 2019); as is a formal notion of 'race' (see Benthall and Haynes 2019). An institutional proposal is in (Veale and Binns 2017).

2.3 Interaction with Machines

Human robot interaction (HRI) and generally the interaction with machines are now academic fields in their own right. There is now significant attention to ethical matters in HRI, the dynamics of perception from both sides, the different interests and the intricacy of the social context, including co-working with robots (e.g. Arnold and Scheutz 2017).

2.3.1 Deception & Authenticity

While AI, with robots or without, can be used to manipulate humans into believing and doing things, it can also be used to drive robots that are problematic since they involve deception, or perhaps violate human dignity or the Kantian requirement of 'respect for humanity' (Lin, Abney, and Jenkins 2017). Humans very easily attribute mental properties to objects, and empathise with them, especially when their outer appearance is similar to living beings. This can be used to deceive humans (or animals) into attributing more intellectual or even emotional significance to robots or AI systems than they deserve. For this reason, some parts of humanoid robotics are problematic and there are cases, with a focus on being as lifelike as possible, that have been clearly deceptive for public-relations purposes (e.g. Hanson Robotics' "Sophia"), and others that hover on the edge (e.g. Hiroshi Ishiguro's remote-controlled Geminoids). At times users are made to believe they interact with a person, where they are not. However, some fairly basic constraints of business ethics and law apply to robots, too: product safety and liability, or non-deception in advertisement. It appears that these existing laws take care of many 'concerns' that are raised. There are cases, however, where human-human interaction has aspects that appear specifically human in ways that can perhaps not be replaced by robots, even if deception is not at play: care, love and sex.

2.3.2 Example a) Care Robots

The use of robots in health care for humans is currently at the level of concept studies in real environments, but it may become a usable technology in a few years, and has raised a number of concerns for a dystopian future of de-humanised care (A. Sharkey and Sharkey 2011; Robert Sparrow 2016). This involves robots that support human carers [caregivers] e.g. in lifting patients, or transporting material, robots that enable patients to do certain things by themselves (e.g. eat with a robotic arm), but also robots that are given to patients as company and comfort (e.g. the 'Paro' robot seal). For an overview, see (van Wynsberghe 2016; Nørskov 2017; Fosch-Villaronga and Albo-Canals 2019), for a survey of users (Draper et al. 2014).

One reason why the issue of care has come to the fore is that people have argued we need (humanoid) robots in ageing societies to care for the elderly. This argument makes dubious assumptions, namely that with longer lifespan people will need more care, and that it will not be possible to attract more humans to caring professions. It may also show a bias about age (Jecker forthcoming). Most importantly, it ignores the nature of automation, which is not simply about replacing humans, but about allowing humans to work more efficiently. In fact much of the use of robots in people who need care will not be to support or replace carers, but to enable such people to perform the actions for which they needed others, e.g. walk, eat, wash, or just pick up something from the floor.

It is not very clear that there really is an issue here, since the discussion mostly focuses on the fear of robots de-humanising care, but the actual and foreseeable robots in care are for classic automation of technical tasks as assistive robots (lifting patients, carrying medicine and supplies). They are thus 'care robots' in a behavioural sense of doing what is required, not in sense that a human cares for the patients and the 'cared for' recognises this intention. It appears that the success of 'being cared for' relies on this intentional sense of 'care', which foreseeable robots cannot provide. If anything, the risk of robots in care is the absence of care – because human carers may be needed less. Interestingly, caring for something, even a virtual agent, can be good for the carer themselves (Lee et al. 2019).

A system that pretends to care would be deceptive and thus problematic – unless the deception is countered by sufficiently large utility gain (Coeckelbergh 2016). Some robots that pretend to 'care' on a basic level are available (Paro seal) and others are in the making. There is a question whether these should always be worse than a bad situation in a care home. Perhaps feeling cared for by a machine, to some extent, can be progress in some cases – but in those cases should we go for the 'robotic' path of progress?

2.3.3 Example b) Sex Robots

It has been argued by several tech optimists that humans will likely be interested in sex and companionship with robots and feel good about it (Levy 2007) – this idea has inspired several movies. Given the variation of human sexual preferences, including sex toys and sex dolls, this seems very likely: The question is whether such devices should be manufactured and promoted, and whether there should be limits to use in this murky area. It has been mostly ignored for a long time, but seems to have moved into the mainstream of 'robot philosophy' in recent times (Sullins 2012; Danaher and McArthur 2017; N. Sharkey et al. 2017; Bendel 2018; Devlin 2018).

Humans have long had deep emotional attachments to objects, so perhaps companionship or even love with a predictable android is attractive to some people, especially those that struggle with actual humans, and already prefer dogs, cats, a bird, a computer or a *tamagotchi*. Danaher (forthcoming-b) argues against (Nyholm and Frank 2017) that this can be true friendship, and thus is a valuable goal. It certainly looks like such friendship might increase overall utility, even if lacking in depth. In all this area there is an issue of deception, since a robot cannot (at present) mean what it says, or have feelings for a human – so any expressions of this sort may be deceptive for some users. It is well known that humans are prone to attribute feelings and thoughts to entities that behave as if they had sentience, and even to clearly inanimate objects that show no behaviour at all. Also, paying for deception seems to be an elementary part of the traditional sex industry.

Finally, there are concerns that have often accompanied matters of sex, namely consent (Frank and Nyholm 2017), aesthetic concerns and the worry that humans may be 'corrupted' by certain experiences. Old fashioned though this may seem, human behaviour is influenced by experience, and it is likely that pornography or sex robots support the perception of other humans as mere objects of desire, or even as recipients of abuse, and thus ruin a deeper sexual and erotic experience. The 'Campaign Against Sex Robots' argues that these devices are a continuation of slavery and prostitution and further abuse of humans (Richardson 2017).

2.4 The Effects of Automation on Employment

Industrial automation typically means that individual productivity of human workers is increased, and thus fewer humans are required for the same output. Classic automation replaced human muscle; digital automation replaces human thought or information-processing – and unlike physical machines digital automation is cheap to duplicate (Bostrom and Yudkovski 2014). Automation does not necessarily mean a loss of overall employment, however, because demand can also increase because the available wealth increased. The attempt to increase productivity has probably always been a feature of the economy, though the emphasis on 'growth' is a modern phenomenon (Harari 2016, 240). In the long run, higher productivity has led to more wealth and more employment overall. Major labour market disruptions have occurred in the past, e.g. farming employed over 60% of the workforce in Europe and North-America in 1800, while by 2010 it employed ca. 5% in the EU, even less in the wealthiest countries (Anonymous 2013). Between 1950 and 1970 the number of hired agricultural workers in the UK halved (Zayed and Loft 2019).

In the meantime new jobs and new kinds of jobs have been created and overall wealth increased. The main question is: Is it different, this time? Will the creation of new jobs keep up with the destruction? And even if it is not different, what are the transition costs, and who bears them? Do we need to make societal adjustments for a fair distribution of costs and benefits of automation?

Responses to the issue of unemployment from AI have ranged from the alarmed (Frey and Osborne 2013; Westlake 2014) to the neutral (Metcalf, Keller, and Boyd 2016; Calo 2018) and the optimistic (Brynjolfsson and McAfee 2016; Harari 2016; Danaher forthcoming-a) – though there is agreement that very significant disruption is likely. In principle, the labour market effect of automation seems to be fairly well understood as involving two channels: "(i) the nature of interactions between differently skilled workers and new technologies affecting labour demand and (ii) the equilibrium effects of technological progress through consequent changes in labour supply and product markets." (Goos 2018, 362). And what currently seems to happen in the labour market as a result of AI & robotics automation of information skills is 'job polarisation' or the 'dumbbell' shape (Goos, Manning, and Salomons 2009): The highly skilled technical jobs are in demand and highly paid, the low skilled service jobs are in demand and badly paid, but the mid-qualification jobs in factories and offices, i.e. the majority of jobs, are under pressure and reduced because they are relatively predictable, and most likely to be automated (Baldwin 2019). Whether AI will ultimately drive unemployment will depend both on technical development and on the societal response.

It is not surprising that new technologies will be used to continue unfair economic practices that are available, e.g. excessively cheap human labour in the 'crowd' (Whittaker et al. 2018, 34ff) or other forms of classical capitalist exploitation. – On the other hand, perhaps due to enormous productivity gains, the 'age of leisure' can at last be realised, as (Keynes 1930) had predicted to occur around 2030 at a growth rate of 1% per annum? Actually, we have reached the level he anticipated for 2030, but we are still working – and consuming more. Harari explains how this economical development allowed humanity to overcome hunger disease and war – and now we aim for immortality and eternal bliss through AI (Harari 2016, 75 etc.), a project that he thinks is doomed to lead to the irrelevance of humanity.

In general terms, this issue concerns how goods in a society should be justly distributed, or 'distributive justice'. A standard view is that the principles of justice should be rationally decided from behind a 'veil of ignorance' (Rawls 1971), i.e. as if one does not know what position in a society one would actually take (labourer or industrialist, etc.). What follows from this is less obvious, though Rawls thought the chosen principles would support basic liberties and a distribution that is of greatest benefit to the least-advantaged members of society. In any case, some kind of fair distribution is envisaged that goes beyond just accepting a distribution that was produced by the current socioeconomic forces. It would appear that the AI economy has two features that contradict such justice: First, it operates in a largely unregulated environment where responsibility is often hard to allocate (which may deteriorate into Hobbes' "state of nature"). Second, it operates in markets that have a 'winner takes all' feature where monopolies develop quickly. This second feature harks back to the first, since monopolies are typically regulated.

The question thus is how the additional wealth is distributed, and whether this can be left to free market forces. What are the societal factors and agents that need to be involved? – Again, a question that has been asked with other forms of automation. It is made harder by the move to a 'new economy' with the digital service industry, especially platforms like Google or Facebook, based on intangible assets, also called 'capitalism without capital' (Haskel and Westlake 2017). This has many challenges, of which unemployment is just one: Perhaps more prominent are monopolies and the difficulty of traditional policy to control multinational digital corporations that do not rely on physical plant in a particular location. Generally, the benefits and disadvantages of AI & robotics technologies will be broad and we will need to see in each area how to fairly distribute these, e.g. influences on the natural environment, on the workplace, on landscapes and cityscapes, on healthcare, on information, etc. etc.

2.5 Autonomous Systems

2.5.1 Autonomy Generally

There are several notions of autonomy in the discussion. A stronger notion of self-governing 'autonomy' is involved in philosophical discussions, where autonomy is the basis for responsibility and personhood (Christman 2018). In this context, responsibility implies autonomy, but responsibility is the more narrow and more demanding notion, so there can be systems that have degrees of technical autonomy without raising issues of responsibility. While some systems will raise issues of responsibility due to increased autonomy, there may be reasons other than responsibility for raising these issues, e.g. the absence of an identifiable individual agent that is stable over time. The weaker, more technical, notion of autonomy in robotics is relative and gradual: A system is said to be autonomous with respect to human control to a certain degree (Müller 2012). There is a parallel here to the issues of bias and opacity in AI since autonomy also concerns a power-relation: who is in control, and who is responsible. It is not just the user, so who or what is it?

Generally speaking, one question is whether autonomous robots raise issues that suggest a revision of present conceptual schemes, or whether they just require technical adjustments to what we already have. In most jurisdictions, there is a sophisticated system of civil and criminal liability, in particular product liability, which may resolve some of the issues. Technical standards, e.g. for the safe use of machinery in medical environments, or for product audits and other forms of regulatory compliance, will likely need to be adjusted, but perhaps not overthrown by fundamental problems. There is already a field of 'verifiable AI' for such safety-critical systems, and for 'security applications'. Bodies like the IEEE and the BSI have produced 'standards', particularly on more technical sub-problems, such as data security and transparency. Among the many autonomous systems on land, on water, under water, in the air or in space, we discuss two samples: autonomous vehicles and autonomous weapons.

Autonomous systems are discussed here under the assumption that they are used by humans and the responsibility remains with the users or makers – if the autonomy is strong enough to raise the issue whether the robots themselves are *responsible* for their actions then the matter is discussed in section 3.2 'Artificial Moral Agents'.

2.5.2 Example a) Autonomous Vehicles

Autonomous vehicles hold the promise to reduce the very significant damage that driving currently causes – with approximately 1 million humans being killed per year, many more injured, the environment polluted, earth sealed with concrete and tarmac, cities full of parked cars, etc. etc. However, there seem to be questions on how autonomous vehicles should behave, and how responsibility and risk should be distributed in the complicated system of autonomous vehicle and traffic control. (There is also significant disagreement over how long the development of fully autonomous, or 'level 5', cars will actually take. These would be cars that can replace the human driver entirely (SAE 2015).

In the classic 'trolley problems' (J. J. Thompson 1976; Woollard and Howard-Snyder 2016, section 2) various dilemmas are presented, the simplest version is that of a trolley train on a track that is heading towards five people and will kill them, unless the train is diverted onto a side track, but on that track there is one person, who will be killed if this choice is made. The example goes back to a remark in (Foot 1967, 6), who discusses a number of dilemma cases where tolerated and indented consequences of an action differ. The 'trolley problems' are not supposed to describe actual ethical problems or to be solved with a 'right' choice. They are thought experiments where choice is artificially constrained to a small finite number of distinct one-off options and the agents has perfect knowledge. These problems are used as a theoretical tool to investigate ethical intuitions and theories – especially the difference between actively doing vs. allowing something to happen, intended vs. tolerated consequences, and consequentialist vs. other normative approaches (Kamm and Rakowski 2016). As such they are often used in teaching introductory ethics.

This type of problem has reminded many of the problems encountered in actual driving, and in autonomous driving (Lin 2015). It is doubtful, however, that an actual driver or autonomous car will ever have to solve trolley problems (but see Keeling forthcoming). While autonomous car trolley problems have received a lot of media attention (Awad et al. 2018), they do not seem to offer anything new to either ethical theory or to the programming of autonomous vehicles, beyond a useful 'intuition pump'.

The more common ethical problems in driving, such as speeding, risky overtaking, not keeping a safe distance, not making space for emergency vehicles, etc. etc. are classic problems of pursuing personal interest vs. the common good. The vast majority of these are also covered by legal regulations on driving, so programming the car to drive 'by the rules' rather than 'by the interest of the passengers' or 'to achieve maximum utility' is deflated to a standard problem of programming ethical machines (see section 3.1). There are probably additional discretionary rules of politeness, e.g. on allowing others to join a priority road, allowing large vehicles more space than they officially deserve, or giving way to driving errors by weaker participants in traffic. What might be more interesting is when to break the rules, e.g. to avoid an accident (Lin 2015), but again this seems to be more a case of applying standard considerations for driving (rules vs. utility) to the case of autonomous vehicles.

Notable policy efforts in this field include the report (German Federal Ministry of Transport and Digital Infrastructure 2017), which stresses that *safety* is the primary objective and accountability cannot remain with the 'driver'. Rule 10 states "In the case of automated and connected driving systems, the accountability that was previously the sole preserve of the individual shifts from the motorist to the manufacturers and operators of the technological systems and to the bodies responsible for taking infrastructure, policy and legal decisions." (See below (3.2.1).) The resulting German and EU laws on licensing automated driving are much more restrictive than their US counterparts where 'testing on consumers' is a strategy used by some companies – without informed consent of the consumers or their possible victims.

2.5.3 Example b) Autonomous Weapons

The notion of automated weapons is fairly old: "For example, instead of fielding simple guided missiles or remotely piloted vehicles, we might launch completely autonomous land, sea, and air vehicles capable of complex, far-ranging reconnaissance and attack missions." (DARPA 1983, 1). This proposal was ridiculed as 'fantasy' at the time (Dreyfus, Dreyfus, and Athanasiou 1986, ix), but it is now a reality, at least for more easily identifiable targets (missiles, planes, ships, tanks, etc.), but not for human combatants that must be distinguished from civilians. The main arguments against (lethal) autonomous weapon systems (AWS or LAWS), are that they support extrajudicial killings, take responsibility away from humans, and make wars or killings more likely – for a detailed list of issues see (Lin, Bekey, and Abney 2008, 73-86).

One particular concern is whether autonomous weapons would make wars, or violence short of war more likely. Lowering the hurdle to use such systems (autonomous vehicles, 'fire-and-forget' missiles, or drones loaded with explosives) and reducing the probability of being held accountable would increase the probability of their use. One threat that is easy to imagine be a small drone that searches, identifies, pursues and kills an individual human – or perhaps a type of human. The same threat on a larger scale would be conventional cruise missiles to identify and attack targets autonomously. It is also said that autonomous weapons "could wreak havoc in densely populated areas" (Scientific American 2019). These are the kinds of cases brought forward by the *Campaign to Stop Killer Robots* (https://www.stopkillerrobots.org) and other activist groups. Some seem to be equivalent to saying that autonomous weapons *are* weapons – weapons kill, weapons of mass-destruction kill many people at once, but we still make them in gigantic numbers. Perhaps the arguments are arguments against weapons rather than against autonomy in weapons? On the matter of accountability, autonomous weapons might make identification and prosecution of the responsible agents more difficult – but this is not clear, given the digital records that one can keep, at least in a conventional war. Also the greater precision of autonomous weapons may save civilian lives. The crucial asymmetry where one side can kill with impunity already exists in conventional drone wars with remote controlled weapons (e.g. US in Pakistan).

Another crucial question seems to be whether using autonomous weapons in war would make wars worse, or perhaps reduce human suffering overall and make wars less bad? If robots reduce war crimes and crimes in war, the answer may well be positive and has been used as an argument in favour of these weapons (Arkin 2009; Müller 2016a) but also as an argument against (Amoroso and Tamburrini 2018). Arguably the main threat is not the use of such weapons in conventional warfare, but in asymmetric conflicts or by non-state agents, including criminals.

It has also been said that autonomous weapons cannot conform to International Humanitarian Law, which requires observance of the principles of distinction (between combatants and civilians), proportionality (of force) and military necessity (of force) in military conflict (A. Sharkey 2019). It is true that the distinction between combatants and non-combatants is hard, but the distinction between civilian and military ships is easy – so all this says is that we should not construct and use such weapons if they violate Humanitarian Law. A few concerns have been raised that being killed by autonomous weapons threatens human dignity, but even the defenders of a ban on these weapons seem to say that these are not good arguments "There are other weapons, and other technologies, that also compromise human dignity. Given this, and the ambiguities inherent in the concept, it is wiser to draw on several types of objections in arguments against AWS, and not to rely exclusively on human dignity." (A. Sharkey 2019).

A lot has been made in the military guidance on weapons on keeping humans "in the loop" or "on the loop" – these ways of spelling out 'meaningful control' are discussed in (Santoni de Sio and van den Hoven 2018). There have been discussions about the difficulties of allocating responsibility for the killings of an autonomous weapon, and a 'responsibility gap' has been suggested (esp. Rob Sparrow 2007), meaning that neither the human nor the machine may be responsible. On the other hand, we do not assume that for any event there is someone responsible for that event, and the real issue may well be the distribution of risk (Simpson and Müller 2016). The difficulty of allocating punishment is sometimes called the 'retribution gap' (Danaher 2016a) and it looks like a much harder problem than the responsibility or liability gaps. Risk analysis (Hansson 2013) indicates it is crucial to identify who is *exposed* to risk, who is a potential *beneficiary*, and who takes the *decisions* (Hansson 2018, 1822-1824).

3 Ethics for AI & Robotics Systems

3.1 Machine Ethics

Machine ethics is ethics for machines, for 'ethical machines', for machines as *subjects*, rather than for the human use of machines as *objects* (Floridi and Saunders 2004; Moor 2006; Anderson and Anderson 2011; Wallach and Asaro 2017). It is often not very clear whether this is supposed to cover all of AI ethics of to be a part of it. Some pronouncements sound like they include all matters of human machine design: "machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable" (Anderson and Anderson 2007, 15). Other authors on "ethical considerations in the development of intelligent interactive systems" focus on design, leaving out what humans can do with the machine: "AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency." (Dignum 2018, 1, 2). Finally, some of the discussion in machine ethics makes the very substantial assumption that machines can, in some sense, be ethical agents responsible for their actions, or 'autonomous moral agents' (see van Wynsberghe and Robbins 2019).

The basic idea of machine ethics is now finding its way into actual robotics where the assumption that these machines are artificial moral agents in any substantial sense is usually not made (Winfield et al. 2019). It is sometimes observed that a robot that is programmed to follow ethical rules can very easily be modified to follow unethical rules (Vanderelst and Winfield 2018).

The idea that machine ethics might take the form of 'laws' has famously been investigated by Isaac Asimov, who proposed 'three laws of robotics' (Asimov 1942): "First Law – A robot may not injure a human being or, through inaction, allow a human being to come to harm. Second Law – A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third Law – A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws." Asimov then showed in a number of stories how conflicts of these laws will make it problematic to use them, despite their hierarchical organisation. Actual machine ethics for autonomous decisions in real-life situations would seem to involve a level of general intelligence that is well beyond current ability of AI.

It is not clear that there is a consistent notion of 'machine ethics' since weaker versions are in danger of reducing 'having an ethics' to acting according to a rule or other notions that fall short of 'reflecting on action' or even 'acting ethically'. And stronger notions that move towards artificial moral agents may describe a – currently - empty set.

3.2 Artificial Moral Agents

If one takes machine ethics to concern moral agents, in some substantial sense, then the agents in question can be called 'artificial moral agents'. The question then is how to explain that notion, in detail, and which systems are such agents. We propose that the main questions will be about rights and responsibilities. Having said that, the discussion about artificial entities challenges a number of common notions in ethics and it can be very useful to understand these to abstract from the specific human case (cf. Powers and Ganascia forthcoming).

Several authors use 'artificial moral agent' in a less demanding sense, borrowing from the software 'agent' use (Allen, Varner, and Zinser 2000), in which case matters of responsibility and rights will not arise. James Moor distinguishes four types of machine ethics: ethical impact agents (example: robot jockeys), implicit ethical agents (example: safe autopilot), explicit ethical agents (example: using formal methods to estimate utility), and full ethical agents ("can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent.") The earlier authors (Allen, Smit, and Wallach 2005; Moor 2006) propose several ways to achieve explicit ethical agents, via programming it in (operational morality), via 'developing' the ethics itself (functional morality) and finally full-blown morality with full intelligence and sentience. In some discussions the notion of 'patient' plays a role: Ethical *agents* have responsibilities while ethical *patients* have rights (because harm to them matters). (Torrance 2011) suggests "artificial (or machine) ethics could be defined as designing machines that do things which, when done by humans, are criterial of the possession of 'ethical status' in those humans" – which he takes to be "ethical *productivity* and ethical *receptivity*" – his expressions for agents and patients. It seems clear that some entities are patients without being agents, e.g. simple animals that can feel pain but not make justified choices. On the other hand it is normally understood that agents will also be patients (e.g. in a Kantian framework). Programmed agents may not be considered 'full' agents because they are "competent without comprehension", just like the neurons in a brain (Dennett 2017), meaning they challenge our ethical system of moral agency where agency and responsibility normally go together. It has also been said that there can be no 'machine ethics' because an ethics only deserves the name if the system has autonomously given it to itself and accepted it – so it cannot be 'programmed in' (Weber 2019) – this rejects all four of Moor's types.

3.2.1 Responsibility for Robots

If the robots act, will they themselves be responsible, liable or accountable for their actions, in some sense? Or perhaps the distribution of risk should take precedence over discussions of responsibility? This concerns more substantial notions of autonomy than the technical one used above in (2.4).
There is broad consensus that accountability, liability, and the rule of law are basic requirements that must be upheld in the face of new technologies (European Group on Ethics in Science and New Technologies 2018, 18), but the issue is how can this be done, and how can responsibility be allocated. Traditional distribution of responsibility already occurs: A car maker is responsible for the technical safety of the car, a driver is responsible for driving it safely, the public authorities are responsible for the technical conditions of the roads, traffic lights, etc. In general "The effects of decisions or actions based on AI are often the result of countless interactions among many actors, including designers, developers, users, software, and hardware. … With distributed agency comes distributed responsibility." (Taddeo and Floridi 2018, 751). How this distribution might occur is not a problem that is specific to AI, but it gains particular urgency in this context (Nyholm 2018a), e.g. for autonomous vehicles (Nyholm 2018b). In classical control engineering, distributed control is often achieved through a control hierarchy plus control loops across these hierarchies. (See also the discussion on autonomous systems above.)

3.2.2 Rights for Robots

Some authors have indicated that it should be seriously considered whether even current robots must be allocated rights (Gunkel 2018a, 2018b; Turner 2019). It is not clear that there are positive arguments for this position; it seems to rely largely on criticism of the opposite position and on the empirical observation that robots and other non-persons are sometimes treated as having rights. In this vein, a 'relational turn' has been proposed: If we relate to robots as though they had rights, then this is fine and we might be well-advised not to search whether they 'really' do have such rights (Coeckelbergh 2012, 2018). This raises the question how far such anti-realism or quasi-realism can go, and what it means then to say that 'robots have rights' in a human-centred approach (Gerdes 2016). On the other side of the debate, Bryson has insisted with a useful [but admittedly problematic] slogan, that "robots should be slaves" (Bryson 2008), i.e. not enjoy rights (Gunkel and Bryson 2014). It has also been said that the reasons for developing robots with rights, or artificial moral patients, in the future are ethically doubtful (van Wynsberghe and Robbins 2019) – this issue is easily confused with the issue of whether 'machine ethics' is a useful notion.
Usually, being a person is supposed to be what makes an entity a responsible agent, someone who can have duties and be the object of ethical concerns – such personhood is typically a deep notion associated with free will (Frankfurt 1971) and perhaps all there is to that notion (Strawson 2004), or to the notion of being the kind of thing that can be praised or blamed (thus influencing future behaviour). It is often, but perhaps not necessarily, associated with having phenomenal consciousness. In order to have rights, being a person or a moral *agent* is not necessary – it is sufficient to be a moral *patient* that deserves moral consideration, e.g. an animal that can feel pain.
There is a wholly separate issue whether robots (or other AI systems) should be given the status of 'legal entities', or 'legal persons' – in a sense in which natural persons, but also states, businesses or organisations are 'entities', namely they can have legal rights and duties, so they can enter into legal contracts, can be sued, etc. The European Parliament has considered allocating such status to robots in order to deal with civil liability (Parliament 2016; Bertolini and Aiello 2018), but not criminal liability – which is reserved for natural persons. It would also be possible to assign only a certain subset of rights and duties to robots. It has been said that "such legislative action would be morally unnecessary and legally troublesome" because it would not serve the interest of humans (Bryson, Diamantis, and

Grant 2017, 273). In environmental ethics there is a long-standing discussion about the legal rights for natural objects like trees (C. D. Stone 1972).

In the community of 'artificial consciousness' researchers there is a significant concern whether it would be ethical to create such consciousness, since creating it would presumably imply ethical obligations to a sentient being, e.g. not to harm it and not to end its existence by switching it off – some authors have called for a "moratorium on synthetic phenomenology" (Bentley et al. 2018, 28f).

4 Singularity

4.1 Singularity and Superintelligence

The idea of the singularity is that if the trajectory of artificial intelligence reaches up to systems that have a human level of intelligence, then these systems would themselves seem to have the ability to develop further AI that surpasses human level, that is 'superintelligent'. This sharp turn of events is the 'singularity', from where onwards the development is out of human control. (So the terms 'singularity' and 'superintelligence' do not refer to the same object.) In some parts of the field it is common to distinguish 'technical' or 'narrow' AI from the aim of 'artificial general intelligence' (AGI), which sometimes understood as identical to Searle's notion of 'strong AI': "computers given the right programs can be literally said to *understand* and have other cognitive states" (Searle 1980, 417).

The fear "the robots we created will take over the world" had captured human imagination even before there were computers (e.g. Butler 1863) and it is the central theme in Čapek's famous play that introduced the word 'robot' before actual robots existed (Čapek 1920) – science fiction has always been a strong inspiration for AI. It was first formulated as a possible trajectory of existing AI into an 'intelligence explosion' to superintelligence by Irvin Good: "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, provided that the machine is docile enough to tell us how to keep it under control" (Good 1965, 33).

The optimistic argument from acceleration to singularity is spelled out by Kurzweil (1999, 2005, 2012), who essentially points out that computing power has been increasing exponentially, i.e. doubling ca. every 2 years since 1970 in accordance with 'Moore's Law' on the number of transistors, and will continue to do so for some time in the future. He predicted in (Kurzweil 1999) that by 2010 supercomputers will reach human computation capacity, by 2030 'mind uploading' will be possible, and by 2045 the 'singularity' occurs where computers far surpass humans, future development of intelligence is taken over by AI, and history is changed forever. The version of this argument that is now used more commonly (Chalmers 2010) talks about an increase in 'intelligence' of the AI system (rather than raw computing power), but the crucial point of 'singularity' remains the one where further development of AI is taken over by AI systems and accelerates beyond human level. (Bostrom 2014) explains in some detail what would happen if intelligence clearly surpasses human level, which he calls 'superintelligence', and what the risks for humanity are. The discussion is nicely summarised in (Eden et al. 2012) and (Shanahan 2015). There are other possible paths to superintelligence, e.g. the complete emulation of the human brain on a computer (Kurzweil 2012; Sandberg 2013), biological paths, or networks and organisations (Bostrom 2014, 22-51).

Despite obvious weaknesses with the identification of 'intelligence' with processing power, Kurzweil seems right that humans tend to underestimate the power of exponential growth. Mini-test: If you walked in steps in such a way that each step is double the previous, starting with a step of one metre, how far would you get with 30 steps? (Answer: to Earth's only permanent natural satellite.) Indeed most progress in AI is readily attributable to the availability of degrees of magnitude faster processors and larger storage (Müller 2018). The actual acceleration and its speeds are discussed in (Müller and Bostrom 2016; Bostrom, Dafoe, and Flynn forthcoming); while (Sandberg 2019) argues that progress will continue for some time.

The participants in this debate are united by being technophiles, in the sense that they expect technology to develop rapidly and bring broadly welcome changes – but beyond that, they divide into those that focus on benefits (e.g. Kurzweil) or on risks (e.g. Bostrom). Both camps sympathise with 'transhuman' views of survival for humankind in a different physical form (e.g. uploaded on a computer) in authors like (Moravec 1990, 1998) and (Bostrom 2003a, 2003c). They also consider the prospects of 'human enhancement', in various respects, including intelligence - often called "IA" (intelligence augmentation), rather than AI. The notion of 'human' itself is up for grabs. It may be that future AI will be used for human enhancement, or will contribute further to the dissolution of the neatly defined human single person. Robin Hanson provides detailed speculation about what will happen economically in case human 'brain emulation' enables truly intelligent robots or 'ems' (Hanson 2016). The argument from superintelligence to existential risk requires the assumption that superintelligence does not imply benevolence – contrary to several traditions in ethics, notably Kant (against Hume), that have argued higher levels of rationality or intelligence would go along with a better understanding of what is moral, and better ability to act morally (Gewirth 1978; Chalmers 2010, 36f). Arguments for existential risk from superintelligence typically deny this and say that rationality and morality are entirely independent or "orthogonal" dimensions – this is sometimes explicitly argued for as an "orthogonality thesis" (Bostrom 2012; Armstrong 2013; Bostrom 2014, 105-109). It is also conceivable that a superintelligence could decide to end humans existence for good ethical reasons.

Criticism has been raised from various angles. Kurzweil and Bostrom seem to assume that intelligence is a one-dimensional property and that the set of intelligent agents is well-ordered, in the mathematical sense – but neither discusses intelligence at any length in their books. Generally, it is fair to say that despite some efforts, the assumptions made in the powerful narrative of superintelligence and singularity have not been investigated in detail. One question is whether such a singularity will ever occur – it may be conceptually impossible, practically impossible or just not happen because of contingent events, including people actively preventing it. Philosophically, the interesting question is whether singularity is just a 'myth' (Floridi 2016; Ganascia 2017) that is not on the trajectory of actual AI research; which is something that practitioners, especially from robotics, often support (e.g. Brooks 2017). They may do so because they fear the PR backlash, because they overestimate the practical problems, or because they have good reasons to think that superintelligence is an unlikely outcome of current AI research (Müller forthcoming). This discussion raises the question whether the concern about 'singularity' is just a narrative about fictional AI that exploits human fears. But even if one *does* find negative reasons compelling and the singularity not likely to occur, there is still a significant possibility that one may turn out to be wrong in this view. Philosophy is not on the 'secure path of a science' (Kant 1791, B15), and maybe AI and robotics aren't either (Müller 2019). So, it appears that discussion of the very high-impact risk of singularity has justification *even if* one thinks the probability of such singularity ever occurring is very low – as long as it is not too low (and that border is, again, a matter for discussion).

4.2 Existential Risk from Superintelligence

Thinking about superintelligence in the long term raises the question whether superintelligence may lead to the extinction of the human species, which is called an "existential risk" (or XRisk) for our species: The superintelligent systems may well have preferences that conflict with the existence of humans on Earth, and may thus decide to end that existence – and given their superior intelligence, they will have the power to do so (or they may happen to end it because they do not really care). Perhaps there is even an astronomical pattern that an intelligent species is bound to discover AI at some point, and thus bring about its own demise. Such a 'great filter' would contribute to the explanation of the "Fermi paradox" why there is no sign of life in the known universe despite the high probability of it emerging – and it would be bad news if we found out that the 'great filter' is ahead of us, rather than an obstacle that Earth has already passed. These issues are sometimes taken more narrowly to be about human extinction (Bostrom 2013), or more broadly as concerning any large risk for the species (Rees 2018) – of which AI is only one (Häggström 2016). Bostrom also uses the category of 'global catastrophic risk' for risks that are sufficiently high up the two dimensions of 'scope' and 'severity' (Bostrom and Ćirković 2011; Bostrom 2013). These discussions of risk are usually not connected to the general problem of ethics under risk (e.g. Hansson 2013, 2018).

Thinking in the long term, even on an astronomical scale, is the crucial feature of this literature. Whether the singularity (or another catastrophic event) occurs in 30 or in 300 or 3000 years does not really matter (Baum et al. 2019). This is part of an approach to focus more work on risks that have low probability but high impact, based on the observation that, traditionally, humans seem to spend a disproportionate amount of resources on risks with high probability but low-impact, even though seen from a classical expected utility calculation (decision theory, cost-benefit analysis) it may be more rational to tackle low probability high-impact risks (North 1968). The long-term view has its own methodological challenges, but has produced a wide discussion: (Häggström 2016) provides an overview of high-impact technological risk for humanity; (Tegmark 2017) focuses on AI and human life '3.0' after singularity while (Russell, Dewey, and Tegmark 2015) and (Bostrom, Dafoe, and Flynn forthcoming) survey longer-term policy issues in ethical AI. Several collections of papers have investigated the risks of artificial general intelligence (AGI) and the factors that might make this development more or less risk-laden (Müller 2016b; Callaghan et al. 2017; Yampolskiy 2018), including the development of non-agent AI (Drexler 2019).

4.3 Controlling Superintelligence?

In a narrow sense, the 'control problem' is how we humans can remain in control of an AI system once it is superintelligent. (Should we put it into a 'box' of sorts, should we hard-wire some aspects, should we prevent it from ignoring human intention, …?) (Bostrom 2014, 127ff). In a wider sense it is the problem how we can make sure an AI system will turn out to be positive, in the sense we humans perceive this (at least in the few areas where we do agree). The latter issue is sometimes called 'value alignment' of superintelligent AI; (Hadfield-Menell et al. 2016) define it formally as 'cooperative inverse reinforcement learning'.

How easy or hard it is to control a superintelligence depends to a significant extent on the speed of 'take-off' from a system that is under human control to a superintelligent system. This has led to a particular attention to systems with self-improvement, such as AlphaZero (Silver et al. 2018) that have shown significant improvement in digital games (chess, Go, video-games) over short time periods – though they can currently not improve on their method of improvement.

One aspect of this problem is that we might decide a certain feature is desirable, but then find out that it has unforeseen consequences that are so negative that we would not desire that feature after all. This is the ancient problem of King Midas who wished that all he touches turn into gold (as Stuart Russell pointed out), also captured in the expression "be careful what you wish for" (Armstrong 2014, 28ff). For example, the designer Robert Propst invented the flexible 'Action Office', but what came of it was the dreaded cubicle office furniture system in open plan spaces. The computer 'Hal' in the 1968 film "2001: A Space Odyssey" tries to kill the human crew in order to carry on with the mission – just as it was programmed to do. This problem has been discussed on the occasion of various examples, such as the 'paperclip maximiser' (Bostrom 2003b), or the program to optimise chess performance (Omohundro 2014). Despite orthogonality, superintelligent systems are typically imagined as having goals, including the goal of self-preservation. The problem is similar to the problem of expressing ethics in the form or laws, even with a hierarchy, in that both rely on the hope that the imagination used was sufficient to foresee all problems. There always seems to remain a risk that a well-intended system will have unforeseen negative consequences. Generally, programming means rather than ends is dangerous (Yudkowsky 2008), but even programming ends can lead to undesired outcomes.

These approaches speculate about ultimately incomprehensible omniscient beings, the radical changes in a 'latter day', and the promise of immortality through transcendence of our current bodily form – so they have clear religious undertones (Capurro 1993; O'Connell 2017, 160ff). These issues also pose a well-known problem of epistemology: Can we know the ways of the omniscient? How does one do "AI theology" (Danaher 2015)? The usual opponents have already shown up: The atheists and the nihilists (Gerz 2018). The atheists slogan is "People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world." (Domingos 2015) – in other words, they say we need an ethics for the 'small' problems that occur with actual AI & robotics (sections 2 and 3 above), less for the 'big ethics' of existential risk from AI.

5 Closing

The singularity thus raises the problem of the image of AI again. It is remarkable how imagination or 'vision' has played a central role since the very beginning of the discipline at the 'Dartmouth Summer Research Project' (McCarthy et al. 1955; Simon and Newell 1958). And the evaluation of this vision is subject to dramatic change: In a few decades, we went from the slogans "AI is impossible" (Dreyfus 1972) and "AI is just automation" (Lighthill 1973) to "AI is will solve all problems" (Kurzweil 1999) and "AI may kill us all" (Bostrom 2014). This created media attention and PR efforts, but it also raises the problem how much of this 'philosophy and ethics of AI' is really about AI, rather than about an imagined technology. – As we said at the outset, AI and robotics have raised fundamental questions about what we should do with these systems, what the systems themselves should do, and what risks they have in the long term. They also challenge the human view of humanity as the intelligent and dominant species on Earth. We have seen issues that have been raised and we will have to watch technological and social developments closely to catch the new ethical issues early on, and to develop the necessary philosophical analysis.

6 Bibliography

AI HLEG, 2019, "High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI". *European Commission, 09.04.2019*. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Allen, Colin, Iva Smit, and Wendell Wallach, 2005, “Artificial Morality: Top-Down, Bottom-up, and Hybrid Approaches”. *Ethics and Information Technology, 7* (3), 149-155. https://doi.org/10.1007/s10676-006-0004-4
Allen, Colin, Gary Varner, and Jason Zinser, 2000, “Prolegomena to Any Future Artificial Moral Agent”. *Journal of Experimental & Theoretical Artificial Intelligence, 12* (3), 251-261. http://dx.doi.org/10.1080/09528130050111428
Amoroso, Daniele, and Guglielmo Tamburrini, 2018, “The Ethical and Legal Case against Autonomy in Weapons Systems”. *Global Jurist, 18* (1). https://www.degruyter.com/view/j/gj.2018.18.issue-1/gj-2017-0012/gj-2017-0012.xml
Anderson, Michael, and Susan Leigh Anderson, 2007, “Machine Ethics: Creating an Ethical Intelligent Agent”. *AI Magazine, 28* (4), 15-26. https://doi.org/10.1609/aimag.v28i4.2065
Anderson, Michael, and Susan Leigh Anderson (eds.), 2011, *Machine Ethics*. Cambridge: Cambridge University Press.
Anonymous, 2013, “How Many People Work in Agriculture in the European Union? An Answer Based on Eurostat Data Sources”. *EU Agricultural Economics Briefs, 8*(July 2013). https://ec.europa.eu/agriculture/sites/agriculture/files/rural-area-economics/briefs/pdf/08_en.pdf
Arkin, Ronald C, 2009, *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press.
Armstrong, Stuart, 2013, “General Purpose Intelligence: Arguing the Orthogonality Thesis”. *Analysis and Metaphysics, 12* (68), 1-20.
Armstrong, Stuart, 2014, *Smarter Than Us*. Berkeley: MIRI.
Arnold, Thomas, and Matthias Scheutz, 2017, “Beyond Moral Dilemmas: Exploring the Ethical Landscape in Hri”. *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 445-452.
Asaro, Peter M, 2019, “AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care”. *IEEE Technology and Society Magazine, 38* (2), 40-53. https://ieeexplore.ieee.org/document/8733937
Asimov, Isaac, 1942, “Runaround: A Short Story”. *Astounding Science Fiction, March*, [Reprinted in "I, Robot", New York: Gnome Press 1950, 1940ff].
Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, . . . Iyad Rahwan, 2018, “The Moral Machine Experiment”. *Nature, 563* (7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6
Baldwin, Richard, 2019, *The Globotics Upheaval: Globalisation, Robotics and the Future of Work*. London: Weidenfeld & Nicolson.
Baum, Seth D., Stuart Armstrong, Timoteus Ekenstedt, Olle Häggström, Robin Hanson, Karin Kuhlemann, . . . Roman V. Yampolskiy, 2019, “Long-Term Trajectories of Human Civilization”. *Foresight, 21* (1), 53-83. https://doi.org/10.1108/FS-04-2018-0037
Bendel, Oliver, 2018, “Sexroboter Aus Sicht der Maschinenethik”. In Bendel, Oliver (ed.), *Handbuch Maschinenethik* (pp. 1-19). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-17484-2_22-1
Bennett, Colin J, and Charles Raab, 2006, *The Governance of Privacy: Policy Instruments in Global Perspective* (2nd ed.). Cambridge, Mass.: MIT Press.
Benthall, Sebastian, and Bruce D Haynes, 2019, “Racial Categories in Machine Learning”. *Proceedings of (ACM FAT* ’19), January 29–31, 2019, Atlanta, GA*, 1-10. https://doi.org/10.1145/3287560.3287575
Bentley, Peter J, Miles Brundage, Olle Häggström, and Thomas Metzinger, 2018, “Should We Fear Artificial Intelligence? In-Depth Analysis.”. *European Parliamentary Research Service, Scientific Foresight Unit (STOA), March 2018*(PE 614.547), 1-40. http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA%282018%29614547_EN.pdf
Bertolini, Andrea, and Giuseppe Aiello, 2018, “Robot Companions: A Legal and Ethical Analysis”. *The Information Society, 34* (3), 130-140. https://doi.org/10.1080/01972243.2018.1444249
Binns, Reuben, 2018, “Fairness in Machine Learning: Lessons from Political Philosophy”. *Proceedings of Machine Learning Research, 81* (1), 1-11.
Bostrom, Nick, 2003a, “Are You Living in a Computer Simulation?”. *Philosophical Quarterly, 53* (211), 243-255.
Bostrom, Nick, 2003b, “Ethical Issues in Advanced Artificial Intelligence”. In Smit, I. et al. (ed.), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* (pp. 12-17): Int. Institute of Advanced Studies in Systems Research and Cybernetics. https://nickbostrom.com/ethics/ai.html

Bostrom, Nick, 2003c, “Transhumanist Values”. In Adams, Frederick (ed.), *Ethical Issues for the 21st Century*. Bowling Green: Philosophical Documentation Center Press.

Bostrom, Nick, 2012, “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents”. *Minds and Machines, 22* (2 - special issue ‘Philosophy of AI’ ed. Vincent C. Müller), 71-85. http://link.springer.com/article/10.1007/s11023-012-9281-3

Bostrom, Nick, 2013, “Existential Risk Prevention as Global Priority”. *Global Policy, 4* (1), 15-31.

Bostrom, Nick, 2014, *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, Nick, and Milan M Ćirković (eds.), 2011, *Global Catastrophic Risks*. New York: Oxford University Press.

Bostrom, Nick, Allan Dafoe, and Carrick Flynn, forthcoming, “Policy Desiderata for Superintelligent AI: A Vector Field Approach (V. 4.3)”. In Liao, S Matthew (ed.), *Ethics of Artificial Intelligence*. New York: Oxford University Press. https://nickbostrom.com/papers/aipolicy.pdf

Bostrom, Nick, and Eliezer Yudkovski, 2014, “The Ethics of Artificial Intelligence”. In Frankish, Keith (ed.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge: Cambridge University Press.

Bradshaw, Samantha, Lisa-Maria Neudert, and Phil Howard, 2019, “Government Responses to Malicious Use of Social Media”. *Oxford Project on Computational Propaganda, Working Paper 2019.2*. https://comprop.oii.ox.ac.uk/research/government-responses/

Brooks, Rodney, 2017 (07.09.2017). “The Seven Deadly Sins of Predicting the Future of AI”. https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/

Brownsword, Roger, Eloise Scotford, and Karen Yeung (eds.), 2017, *The Oxford Handbook of Law, Regulation and Technology*. Oxford: Oxford University Press.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, . . . Bobby Filar, 2018, “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation”. *FHI/CSER/CNAS/EFF/OpenAI Report*, 1-101. https://arxiv.org/abs/1802.07228

Brynjolfsson, Erik, and Andrew McAfee, 2016, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W. W. Norton.

Bryson, Joanna J, 2008, “Robots Should Be Slaves”. In Wilks, Yorick (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues* (pp. 63–74). Amsterdam: John Benjamins Publishing.

Bryson, Joanna J, 2019, “The Past Decade and Future of Ai’s Impact on Society”. In Anonymous (ed.), *Towards a New Enlightenment: A Transcendent Decade*. Madrid: Turner - BVVA. https://www.bbvaopenmind.com/en/books/towards-a-new-enlightenment-a-transcendent-decade/

Bryson, Joanna J, Mihailis E Diamantis, and Thomas D Grant, 2017, “Of, for, and by the People: The Legal Lacuna of Synthetic Persons”. *Artificial Intelligence and Law, 25* (3), 273-291. https://doi.org/10.1007/s10506-017-9214-9

Burr, Christopher, and Nello Christianini, forthcoming, “Can Machines Read Our Minds?”. *Minds and Machines*. https://doi.org/10.1007/s11023-019-09497-4

Butler, Samuel, 1863, “Darwin among the Machines: Letter to the Editor”. *The Press (Christchurch), 13.06.1863*. http://nzetc.victoria.ac.nz/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan, 2017, “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases”. *Science, 356* (6334), 183-186. http://science.sciencemag.org/content/sci/356/6334/183.full.pdf

Callaghan, Victor, James Miller, Roman V Yampolskiy, and Stuart Armstrong (eds.), 2017, *The Technological Singularity: Managing the Journey*. Berlin: Springer. https://link.springer.com/book/10.1007/978-3-662-54033-6 - toc

Calo, Ryan, 2018, “Artificial Intelligence Policy: A Primer and Roadmap ”. *University of Bologna Law Review, 3* (2), 180-218.

Calo, Ryan, Michael A Froomkin, and Ian Kerr (eds.), 2016, *Robot Law*. Cheltenham: Edward Elgar.

Čapek, Karel, 1920, *R.U.R.* (Majer, Peter and Cathy Porter, Trans.). London: Methuen 1999.

Capurro, Raphael, 1993, “Ein Grinsen Ohne Katze: Von der Vergleichbarkeit Zwischen 'Künstlicher Intelligenz' und 'Getrennten Intelligenzen'”. *Zeitschrift für philosophische Forschung, 47*, 93-102.

Cave, Stephen, 2019, “To Save Us from a Kafkaesque Future, We Must Democratise AI”. *The Guardian, 04.01.2019*. https://www.theguardian.com/commentisfree/2019/jan/04/future-democratise-ai-artificial-intelligence-power

Chalmers, David J., 2010, “The Singularity: A Philosophical Analysis”. *Journal of Consciousness Studies, 17* (9-10), 7-65. http://consc.net/papers/singularityjcs.pdf

Christman, John, 2018, “Autonomy in Moral and Political Philosophy”. In Zalta, Edward N. (ed.), *Stanford Encyclopedia of Philosophy*. Palo Alto: Stanford University. https://plato.stanford.edu/entries/autonomy-moral/

Coeckelbergh, Mark, 2012, *Growing Moral Relations: Critique of Moral Status Ascription*. London: Palgrave.

Coeckelbergh, Mark, 2016, “Care Robots and the Future of Ict-Mediated Elderly Care: A Response to Doom Scenarios”. *AI & SOCIETY, 31* (4), 455-462.

Coeckelbergh, Mark, 2018, “What Do We Mean by a Relational Ethics? Growing a Relational Approach to the Moral Standing of Plants, Robots and Other Non-Humans”. In Kallhoff, Angela, Marcello Di Paola and Maria Schörgenhumer (eds.), *Plant Ethics* (pp. 110-121). London: Routledge.

Costa, Elisabeth, and David Halpern, 2019, “The Behavioural Science of Online Harm and Manipulation, and What to Do About It: An Exploratory Paper to Spark Ideas and Debate”. *The Behavioural Insights Team Report*, 1-82. https://www.bi.team/publications/the-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it/

Crawford, Kate, and Ryan Calo, 2016, “There Is a Blind Spot in AI Research”. *Nature, 538*, 311-313.

Danaher, John, 2015, “Why AI Doomsayers Are Like Sceptical Theists and Why It Matters”. *Minds and Machines, 25* (3), 231-246. https://philpapers.org/rec/DANTEC-2

Danaher, John, 2016a, “Robots, Law and the Retribution Gap”. *Ethics and Information Technology, 18* (4), 299–309.

Danaher, John, 2016b, “The Threat of Algocracy: Reality, Resistance and Accommodation”. *Philosophy & Technology, 29* (3), 245-268. http://dx.doi.org/10.1007/s13347-015-0211-1

Danaher, John, forthcoming-a, *Automation and Utopia: Human Flourishing in a World without Work*. Cambridge, Mass.: Harvard University Press.

Danaher, John, forthcoming-b, “The Philosophical Case for Robot Friendship”. *Journal of Posthuman Studies*. https://www.researchgate.net/publication/330142494_The_Philosophical_Case_for_Robot_Friendship

Danaher, John, and Neil McArthur (eds.), 2017, *Robot Sex: Social and Ethical Implications*. Boston, Mass.: MIT Press.

DARPA, 1983. “Strategic Computing - New-Generation Computing Technology: A Strategic Plan for Its Development an Application to Critical Problems in Defense (28.10.1983)”. http://www.scribd.com/document/192183614/Strategic-Computing-1983

Dennett, Daniel C, 2017, *From Bacteria to Bach and Back: The Evolution of Minds*. New York: W.W. Norton.

Devlin, Kate, 2018, *Turned On: Science, Sex and Robots*. London: Bloomsbury.

Diakopoulos, Nick, 2015, “Algorithmic Accountability: Journalistic Investigation of Computational Power Structures.”. *Digital Journalism, 3* (3), 398-415.

Dignum, Virginia, 2018, “Ethics in Artificial Intelligence: Introduction to the Special Issue”. *Ethics and Information Technology, 20* (1), 1-3. https://doi.org/10.1007/s10676-018-9450-z

Domingos, Pedro, 2015, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. London: Allen Lane.

Draper, Heather, Tom Sorell, Sandra Bedaf, Dag Sverre Syrdal, Carolina Gutierrez-Ruiz, Alexandre Duclos, and Farshid Amirabdollahian, 2014, “Ethical Dimensions of Human-Robot Interactions in the Care of Older People: Insights from 21 Focus Groups Convened in the UK, France and the Netherlands”. In Beetz, M, B Johnston and MA Williams (eds.), *International Conference on Social Robotics* (Vol. LNCS 8755). Cham: Springer.

Dressel, Julia, and Hany Farid, 2018, “The Accuracy, Fairness, and Limits of Predicting Recidivism”. *Science Advances, 4* (1), 1-5. http://advances.sciencemag.org/content/advances/4/1/eaao5580.full.pdf

Drexler, Eric K, 2019, “Reframing Superintelligence: Comprehensive AI Services as General Intelligence”. *FHI Technical Report, #2019* (1), 1-210. https://www.fhi.ox.ac.uk/research/reports/

Dreyfus, Hubert L., 1972, *What Computers Still Can't Do: A Critique of Artificial Reason* (2 ed.). Cambridge, Mass.: MIT Press 1992.

Dreyfus, Hubert L., Stuart E. Dreyfus, and Tom Athanasiou, 1986, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press.

Eden, Amnon, James H. Moor, Johnny Hartz Søraker, and Eric Steinhart (eds.), 2012, *Singularity Hypotheses: A Scientific and Philosophical Assessment* (*The Frontiers Collection*. Berlin: Springer.

EU Parliament, 2016, “General Data Protection Regulation: Regulation (Eu) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/Ec”. *Official Journal of the European Union, 119*(04.05.2016), 1–88. http://data.europa.eu/eli/reg/2016/679/oj

Eubanks, Virginia, 2018, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. London: St. Martin's Press.

European Group on Ethics in Science and New Technologies, 2018 (09.03.2018). “Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems”. *European Commission, Directorate-General for Research and Innovation, Unit RTD.01.* http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

Ferguson, Andrew Guthrie, 2017, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. New York: NYU Press.

Floridi, Luciano, 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, 09.05.2016*. aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, . . . Effy Vayena, 2018, “Ai4people—an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”. *Minds and Machines, 28* (4), 689-707.

Floridi, Luciano, and Jeff W. Saunders, 2004, “On the Morality of Artificial Agents’”. *Minds and Machines, 14*, 349-379.

Floridi, Luciano, and Mariarosaria Taddeo, 2016, “What Is Data Ethics?”. *Phil. Trans. R. Soc. A, 374* (2083).

Foot, Philippa, 1967, “The Problem of Abortion and the Doctrine of the Double Effect”. *Oxford Review, 5*, 5-15.

Fosch-Villaronga, Eduard, and Jordi Albo-Canals. (2019). "I’ll Take Care of You," Said the Robot *Paladyn, Journal of Behavioral Robotics* (Vol. 10, pp. 77).

Frank, Lily, and Sven Nyholm, 2017, “Robot Sex and Consent: Is Consent to Sex between a Robot and a Human Conceivable, Possible, and Desirable?”. *Artificial Intelligence and Law, 25* (3), 305-323.

Frankfurt, Harry, 1971, “Freedom of the Will and the Concept of a Person”. *The Journal of Philosophy, LXVIII* (1), 5-20.

Frey, Carl Benedikt, and Michael A. Osborne, 2013, “The Future of Employment: How Susceptible Are Jobs to Computerisation?”. *Oxford Martin School Working Papers*. http://www.oxfordmartin.ox.ac.uk/publications/view/1314

Ganascia, Jean-Gabriel, 2017, *Le Mythe De La Singularité*. Paris: Éditions du Seuil.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford, 2018 (23.0.3.2018). “Datasheets for Datasets”. https://arxiv.org/abs/1803.09010

Gerdes, Anne, 2016, “The Issue of Moral Consideration in Robot Ethics”. *SIGCAS Comput. Soc., 45* (3), 274-279.

German Federal Ministry of Transport and Digital Infrastructure, 2017, “Report of the Ethics Commission: Automated and Connected Driving”. (June 2017), 1-36. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html

Gerz, Nolen, 2018, *Nihilism and Technology*. London: Rowman & Littlefield.

Gewirth, Alan, 1978, “The Golden Rule Rationalized”. *Midwest Studies in Philosophy, III* (1), 133-147. https://doi.org/10.1111/j.1475-4975.1978.tb00353.x

Gibert, Martin, 2019, “Éthique Artificielle (Version Grand Public)”. In Kristanek, M (ed.), *Encyclopédie Philosophique*. http://encyclo-philo.fr/etique-artificielle-gp/

Giubilini, Alberto, and Julian Savulescu, 2018, “The Artificial Moral Advisor: the “Ideal Observer” Meets Artificial Intelligence”. *Philosophy & Technology, 31* (2), 169-188. https://doi.org/10.1007/s13347-017-0285-z

Good, Irvin J, 1965, “Speculations Concerning the First Ultraintelligent Machine”. In Alt, Franz L and Morris Ruminoff (eds.), *Advances in Computers* (Vol. 6, pp. 31-88). New York & London: Academic Press.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, 2016, *Deep Learning*. Cambridge, Mass.: MIT Press.

Goodman, Bryce, and Seth Flaxman, 2016, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"". *ARXIV, 06/2016*(1606.08813). https://arxiv.org/abs/1606.08813
Goos, Maarten, 2018, "The Impact of Technological Progress on Labour Markets: Policy Challenges". *Oxford Review of Economic Policy, 34* (3), 362–375.
Goos, Maarten, Alan Manning, and Anna Salomons, 2009, "Job Polarization in Europe". *American Economic Review, 99* (2), 58-63.
Graham, Sandra, and Brian S. Lowery, 2004, "Priming Unconscious Racial Stereotypes About Adolescent Offenders". *Law and Human Behavior, 28* (5), 483-504. https://doi.org/10.1023/B:LAHU.0000046430.65485.1f
Gunkel, David J, 2018a, "The Other Question: Can and Should Robots Have Rights?". *Ethics and Information Technology, 20* (2), 87–99. https://doi.org/10.1007/s10676-017-9442-4
Gunkel, David J, 2018b, *Robot Rights*. Boston, Mass.: MIT Press.
Gunkel, David J, and Joanna J Bryson (eds.), 2014, *Special Issue on Machine Morality* (*Philosophy & Technology,* Vol. 27).
Gunning, David, 2018. "Explainable Artificial Intelligence (Xai)". *Defense Advanced Research Projects Agency.* https://www.darpa.mil/program/explainable-artificial-intelligence
Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell, 2016, "Cooperative Inverse Reinforcement Learning". *NIPS 2016*, 1-9. https://arxiv.org/abs/1606.03137
Häggström, Olle, 2016, *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford: Oxford University Press.
Hanson, Robin, 2016, *The Age of Em: Work, Love and Life When Robots Rule the Earth*. Oxford: Oxford University Press.
Hansson, Sven Ove, 2013, *The Ethics of Risk: Ethical Analysis in an Uncertain World*. New York: Palgrave Macmillan.
Hansson, Sven Ove, 2018, "How to Perform an Ethical Risk Analysis (Era)". *Risk Analysis, 38* (9), 1820-1829.
Harari, Yuval Noah, 2016, *Homo Deus: A Brief History of Tomorrow*. New York: Harper.
Harris, Tristan, 2016, "How Technology Is Hijacking Your Mind—from a Magician and Google Design Ethicist". *medium.com, Thrive Global* (18.05.2016). https://medium.com/thrive-global/how-technology-hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3
Haskel, Jonathan, and Stian Westlake, 2017, *Capitalism without Capital: The Rise of the Intangible Economy*. Princeton, NJ: Princeton University Press.
Houkes, Wybo, and Pieter E Vermaas, 2010, *Technical Functions: On the Use and Design of Artefacts*. Berlin: Springer.
IEEE, Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019, "Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems". *25.03.2019* (1st ed.), 1-294. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html
Jacobs, An, Lynn Tytgat, Michel Maus, Romain Meeusen, and Bram Vanderborght (eds.), 2019, *Homo Roboticus: 30 Questions and Answers on Man, Technology, Science & Art*. Brussels: ASP. http://homo-roboticus.be
Jasanoff, Sheila, 2016, *The Ethics of Invention: Technology and the Human Future*. New York: Norton.
Jecker, Nancy S, forthcoming, *Ending Midlife Bias: New Values for Old Age*. New York: Oxford University Press.
Johnson, Deborah G, and Mario Verdicchio, 2017, "Reframing AI Discourse". *Minds and Machines, 27* (4), 575–590. https://link.springer.com/article/10.1007/s11023-017-9417-6 - citeas
Kahnemann, Daniel, 2011, *Thinking Fast and Slow*. London: Macmillan.
Kamm, Frances Myrna, and Eric Rakowski (eds.), 2016, *The Trolley Problem Mysteries*. New York: Oxford University Press.
Kant, Immanuel, 1791, *Critique of Pure Reason* (Smith, Norman Kemp, Trans.). London: Palgrave Macmillan 1929.
Keeling, Geoff, forthcoming, "Why Trolley Problems Matter for the Ethics of Automated Vehicles". *Science and Engineering Ethics, Online First 04.03.2019*. https://link.springer.com/article/10.1007/s11948-019-00096-1
Keynes, John Maynard, 1930, "Economic Possibilities for Our Grandchildren" *Essays in Persuasion* (pp. 358-373). New York: Harcourt Brace 1932.

Kissinger, Henry 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/

Kurzweil, Ray, 1999, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. London: Penguin.

Kurzweil, Ray, 2005, *The Singularity Is Near: When Humans Transcend Biology*. London: Viking.

Kurzweil, Ray, 2012, *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Viking.

Lee, Minha, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn, 2019, "Caring for Vincent: A Chatbot for Self-Compassion". *(CHI '19) Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (702), 1-13. https://dl.acm.org/citation.cfm?doid=3290605.3300932

Levy, David, 2007, *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper & Co.

Lighthill, James, 1973, "Artificial Intelligence: A General Survey". *Artificial intelligence: A paper symposion,* (London). http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm

Lin, Patrick, 2015, "Why Ethics Matters for Autonomous Cars". In Maurer, M. et al. (ed.), *Autonomous Driving* (pp. 69-85). Berlin: Springer. DOI 10.1007/978-3-662-48847-8_4

Lin, Patrick, Keith Abney, and Ryan Jenkins (eds.), 2017, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press.

Lin, Patrick, George Bekey, and Keith Abney, 2008, "Autonomous Military Robotics: Risk, Ethics, and Design". *US Department of Navy, Office of Naval Research* (December 20, 2008), 1-112. http://ethics.calpoly.edu/ONR_report.pdf

Lomas, Meghann, Robert Chevalier, Ernest Vincent II Cross, Robert Christopher Garrett, John Hoare, and Michael Kopack, 2012, "Explaining Robot Actions" *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 187-188): ACM.

Macnish, Kevin, 2017, *The Ethics of Surveillance: An Introduction*. London: Routledge.

Marcus, Gary, 2018 (02.01.2018). "Deep Learning: A Critical Appraisal". *arXiv.* https://arxiv.org/abs/1801.00631

Mathur, Arunesh, Gunes Acar, Michael Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan, 2019, "Dark Patterns at Scale: Findings from a Crawl of 11k Shopping Websites". *Proceedings of the ACM Human-Computer Interaction, 3* (81), 1-32. https://arxiv.org/abs/1907.07032

McCarthy, John, Marvin Minsky, Nathaniel Rochester, and Claude E. Shannon, 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence". Retrieved October 2006, http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Metcalf, Jacob, Emily F. Keller, and Danah Boyd, 2016, "Perspectives on Big Data, Ethics, and Society". *Council for Big Data, Ethics, and Society, May 23, 2016*, 23pp. http://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/

Minsky, Marvin, 1985, *The Society of Mind*. New York: Simon & Schuster.

Mittelstadt, Brent Daniel, and Luciano Floridi, 2016, "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts". *Science and Engineering Ethics, 22* (2), 303-341.

Moor, James H., 2006, "The Nature, Importance, and Difficulty of Machine Ethics". *IEEE Intelligent Systems, 21* (4), 18-21.

Moravec, Hans, 1990, *Mind Children*. Cambridge, Mass.: Harvard University Press.

Moravec, Hans, 1998, *Robot: Mere Machine to Transcendent Mind*. New York: Oxford University Press.

Mozorov, Eygeny, 2013, *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Müller, Vincent C., 2012, "Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction". *Cognitive Computation, 4* (3), 212-215. http://doi.org/10.1007/s12559-012-9129-4

Müller, Vincent C., 2016a, "Autonomous Killer Robots Are Probably Good News". In Di Nucci, Ezio and Filippo Santoni de Sio (eds.), *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons* (pp. 67-81). London: Ashgate. http://www.ashgate.com/isbn/9781472456724

Müller, Vincent C. (ed.), 2016b, *Risks of Artificial Intelligence*. London: Chapman & Hall - CRC Press. https://www.crcpress.com/Risks-of-Artificial-Intelligence/Muller/9781498734820

Müller, Vincent C., 2018, “In 30 Schritten Zum Mond? Zukünftiger Fortschritt in der Ki”. *Medienkorrespondenz, 20* (05.10.2018), 5-15. https://www.medienkorrespondenz.de
Müller, Vincent C., 2019, “Measuring Progress in Robotics: Benchmarking and the ‘Measure-Target Confusion’”. In Bonsignorio, Fabio, John Hallam, Elena Messina and Angel P Del Pobil (eds.), *Metrics of Sensory Motor Coordination and Integration in Robots and Animals* (pp. 169-179). Berlin: Springer Nature. https://www.springer.com/gp/book/9783030141240
Müller, Vincent C., forthcoming, *Can Machines Think? Fundamental Problems of Artificial Intelligence*. New York: Oxford University Press.
Müller, Vincent C., and Nick Bostrom, 2016, “Future Progress in Artificial Intelligence: A Survey of Expert Opinion”. In Müller, Vincent C. (ed.), *Fundamental Issues of Artificial Intelligence* (pp. 553-570). Berlin: Springer. http://www.springer.com/gp/book/9783319264837
Newport, Cal, 2019, *Digital Minimalism: On Living Better with Less Technology*. London: Penguin.
Nørskov, Marco (ed.), 2017, *Social Robots*. London: Routledge.
North, Warner D, 1968, “A Tutorial Introduction to Decision Theory”. *IEEE Transactions on Systems Science and Cybernetics, 4* (3), 200-210. https://stat.duke.edu/~scs/Courses/STAT102/DecisionTheoryTutorial.pdf
Nyholm, Sven, 2018a, “Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci”. *Science and Engineering Ethics, 24* (4), 1201-1219. https://doi.org/10.1007/s11948-017-9943-x
Nyholm, Sven, 2018b, “The Ethics of Crashes with Self-Driving Cars: A Roadmap, II”. *Philosophy Compass, 13* (7), e12506. https://onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12506
Nyholm, Sven, and Lily Frank, 2017, “From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?”. In Danaher, John and Neil McArthur (eds.), *Robot Sex: Social and Ethical Implications* (pp. 219-243). Cambridge, Mass.: MIT Press.
O'Connell, Mark, 2017, *To Be a Machine: Adventures among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death*. London: Granta.
O’Neil, Cathy, 2016, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Largo, ML: Crown.
Omohundro, Steve, 2014, “Autonomous Technology and the Greater Human Good”. *Journal of Experimental and Theoretical Artificial Intelligence, 26* (3 - Special issue ‘Risks of General Artificial Intelligence’, ed. V. Müller), 303-315.
Parliament, EU, 2016 (31.05.2016). “Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(Inl))”. *Committee on Legal Affairs.*
Powers, Thomas M, and Jean-Gabriel Ganascia, forthcoming, “The Ethics of the Ethics of AI”. In Dubber, Markus D, Frank Pasquale and Sunnit Das (eds.), *Oxford Handbook of Ethics of Artificial Intelligence*. New York. https://c4ejournal.net/the-oxford-handbook-of-ethics-of-ai-online-companion/
Programs, Office of Justice, 2014 (13.01.2014). “Predictive Policing”. *National Institute of Justice.* Retrieved 13.12.2018, https://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/Pages/welcome.aspx
Rawls, John, 1971, *A Theory of Justice*. Cambridge, Mass.: Belknap Press.
Rees, Martin, 2018, *On the Future: Prospects for Humanity*. Princeton: Princeton University Press.
Richardson, Kathleen, 2017, “Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines”. *IEEE Technology and Society* (June 29th, 2017).
Roessler, Beate, 2017, “Privacy as a Human Right”. *Proceedings of the Aristotelian Society, 2* (CXVII).
Royakkers, Lambèr, and Rinie van Est, 2016, *Just Ordinary Robots: Automation from Love to War*. Boca Raton: CRC Press, Taylor & Francis.
Russell, Stuart, Daniel Dewey, and Max Tegmark, 2015, “Research Priorities for Robust and Beneficial Artificial Intelligence”. *AI Magazine, 36* (4), 105-114. http://futureoflife.org/static/data/documents/research_priorities.pdf
SAE, 2015, “Taxonomy and Definitions for Terms Related to Driving Automation Systems for on-Road Motor Vehicles”. *SAE Recommended Practice, J3016_201806* (2018-06-15). https://www.sae.org/standards/content/j3016_201806/
Sandberg, Anders, 2013, “Feasibility of Whole Brain Emulation”. In Müller, Vincent C. (ed.), *Theory and Philosophy of Artificial Intelligence* (pp. 251-264). Berlin: Springer. http://www.pt-ai.org
Sandberg, Anders, 2019, “There Is Plenty of Time at the Bottom: The Economics, Risk and Ethics of Time Compression”. *Foresight, 21*(1), 84-99. https://www.emeraldinsight.com/doi/full/10.1108/FS-04-2018-0044

Santoni de Sio, Filippo, and Jeroen van den Hoven, 2018, "Meaningful Human Control over Autonomous Systems: A Philosophical Account". *Frontiers in Robotics and AI, 5* (15). https://www.frontiersin.org/article/10.3389/frobt.2018.00015

Schneier, Bruce, 2015, *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. New York: W. W. Norton.

Scientific American, The Editors, 2019, "Don't Let Robots Pull the Trigger: Weapons That Kill Enemies on Their Own Threaten Civilians and Soldiers Alike". *Scientific American, 320*(3). https://www.scientificamerican.com/article/dont-let-robots-pull-the-trigger/

Searle, John R., 1980, "Minds, Brains and Programs". *Behavioral and Brain Sciences, 3*, 417-457.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi, 2019, "Fairness and Abstraction in Sociotechnical Systems" *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68). Atlanta, GA, USA: ACM. https://dl.acm.org/citation.cfm?id=3287598

Sennett, Richard, 2018, *Building and Dwelling: Ethics for the City*. London: Allen Lane.

Shanahan, Murray, 2015, *The Technological Singularity*. Cambridge, Mass.: MIT Press.

Sharkey, Amanda, 2019, "Autonomous Weapons Systems, Killer Robots and Human Dignity". *Ethics and Information Technology, 21* (2), 75-87. https://doi.org/10.1007/s10676-018-9494-0

Sharkey, Amanda, and Noel Sharkey, 2011, "The Rights and Wrongs of Robot Care". In Lin, Patrick, Keith Abney and George Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 267-282). Cambridge, Mass.: MIT Press.

Sharkey, Noel, Aimee van Wynsberghe, Scott Robbins, and Eleanor Hancock, 2017, "Report: Our Sexual Future with Robots". *Responsible Robotics*, 1-44. https://responsiblerobotics.org/2017/07/05/frr-report-our-sexual-future-with-robots/

Shoham, Yoav, Perrault Raymond, Brynjolfsson Erik, Jack Clark, James Manyika, Juan Carlos Niebles, . . . Zoe Bauer, 2018 (December 2018). "The AI Index 2018 Annual Report". *AI Index Steering Committee, Human-Centered AI Initiative.* http://cdn.aiindex.org/2018/AI Index 2018 Annual Report.pdf

SIENNA, 2019, "Deliverable Report D 4.4: Ethical Issues in Artificial Intelligence and Robotics". (June 2019), 1-103. http://www.sienna-project.eu

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, . . . Demis Hassabis, 2018, "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play". *Science, 362* (6419), 1140-1144. http://science.sciencemag.org/content/sci/362/6419/1140.full.pdf

Simon, Herbert, and Allen Newell, 1958, "Heuristic Problem Solving: The Next Advance in Operations Research". *Operations Research, 6* (1), 1-10.

Simpson, Thomas W, and Vincent C. Müller, 2016, "Just War and Robots' Killings". *The Philosophical Quarterly, 66* (263), 302-322. https://doi.org/10.1093/pq/pqv075
http://pq.oxfordjournals.org/content/66/263/302.abstract

Smolan, Sandy, 2016, "The Human Face of Big Data". *PBS Documentary,* (24 February 2016), 56 mins. https://www.youtube.com/watch?v=kAZ8lK224Kw

Sparrow, Rob, 2007, "Killer Robots". *Journal of Applied Philosophy, 24* (1), 62-77.

Sparrow, Robert, 2016, "Robots in Aged Care: A Dystopian Future". *AI & SOCIETY, 31* (4), 1-10.

Stahl, Bernd Carsten, Job Timmermans, and Brent Daniel Mittelstadt, 2016, "The Ethics of Computing: A Survey of the Computing-Oriented Literature". *ACM Computing Surveys, 48/4* (55), 1-38.

Stahl, Bernd Carsten, and David Wright, 2018, "Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation". *IEEE Security & Privacy, 16* (3).

Stone, Christopher D, 1972, "Should Trees Have Standing - toward Legal Rights for Natural Objects". *Southern California Law Review* (2), 450-501.

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, . . . Astro Teller, 2016. "Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel". https://ai100.stanford.edu/2016-report

Strawson, Galen, 2004 (29.02.2004). "Free Will". *Routledge Encyclopedia of Philosophy.* Retrieved May 2005, 2005, http://www.rep.routledge.com/article/V014

Sullins, John P, 2012, "Robots, Love, and Sex: The Ethics of Building a Love Machine". *IEEE Transactions on Affective Computing, 3* (4), 398-409.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum, 2019, "Technology, Autonomy, and Manipulation". *Internet Policy Review, 8* (2). https://policyreview.info/articles/analysis/technology-autonomy-and-manipulation

Taddeo, Mariarosaria, and Luciano Floridi, 2018, "How AI Can Be a Force for Good". *Science, 361* (6404), 751-752. http://science.sciencemag.org/content/sci/361/6404/751.full.pdf
Taylor, Steve, Brian Pickering, Michael Boniface, Michael Anderson, David Danks, Asbjørn Følstad, . . . Fiona Wollard, 2018 (June 2018). "Responsible AI – Key Themes, Concerns & Recommendations for European Research and Innovation: Summary of Consultation with Multidisciplinary Experts". https://www.hub4ngi.eu
Tegmark, Max, 2017, *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf.
Thompson, Judith Jarvis, 1976, "Killing, Letting die and the Trolley Problem". *Monist, 59*, 204-217.
Thompson, Nicholas, and Ian Bremmer, 2018, "The AI Cold War That Threatens Us All". *Wired,* (23.10.2018). https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/
Torrance, Steve, 2011, "Machine Ethics and the Idea of a More-Than-Human Moral World". In Anderson, Michael and Susan Leigh Anderson (eds.), *Machine Ethics* (pp. 115-137). Cambridge: Cambridge University Press.
Trump, Donald J, 2019, "Executive Order on Maintaining American Leadership in Artificial Intelligence". *The White House* (11.02.2019). https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/
Turner, Jacob, 2019, *Robot Rules: Regulating Artificial Intelligence*. Berlin: Springer.
Tzafestas, Spyros G, 2016, *Roboethics: A Navigating Overview*. Berlin: Springer.
UK Equality Act, 2010. " https://www.legislation.gov.uk/ukpga/2010/15/contents
Vallor, Shannon, 2017, *Technology and the Virtues: A Philosophical Guide for a Future Worth Wanting*. New York: Oxford University Press.
Van Lent, Michael, William Fisher, and Michael Mancuso, 1999, "An Explainable Artificial Intelligence System for Small-Unit Tactical Behavior" *Proceedings of the National Conference on Artificial Intelligence* (pp. 900-907). Menlo Park, CA: AAAI Press.
van Wynsberghe, Aimee, 2016, *Healthcare Robots: Ethics, Design and Implementation*. London: Routledge.
van Wynsberghe, Aimee, and Scott Robbins, 2019, "Critiquing the Reasons for Making Artificial Moral Agents". *Science and Engineering Ethics, 25* (3), 719-735. https://doi.org/10.1007/s11948-018-0030-8
Vanderelst, Dieter, and Alan Winfield, 2018, "The Dark Side of Ethical Robots". *AAAI/ACM Conference on AI Ethics and Society, 2018*, 1-6. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_98.pdf
Veale, Michael, and Reuben Binns, 2017, "Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data". *Big Data & Society, 4* (2), 2053951717743530. https://doi.org/10.1177/2053951717743530
Verbeek, Peter-Paul, 2011, *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.
Wachter, Sandra, and Brent Daniel Mittelstadt, forthcoming, "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI". *Columbia Business Law Review*, (September 13, 2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248829
Wachter, Sandra, Brent Daniel Mittelstadt, and Luciano Floridi, 2017, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". *International Data Privacy Law*. http://dx.doi.org/10.2139/ssrn.2903469
Wachter, Sandra, Brent Daniel Mittelstadt, and Chris Russell, 2018, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the Gdpr". *Harvard Journal of Law & Technology, 31* (2). http://dx.doi.org/10.2139/ssrn.3063289
Wallach, Wendell, and Peter M Asaro (eds.), 2017, *Machine Ethics and Robot Ethics*. London: Routledge.
Walsh, Toby, 2018, *Machines That Think: The Future of Artificial Intelligence*. Amherst, Mass.: Prometheus Books.
Weber, Karsten, 2019, "Autonomie und Moralität Als Zuschreibung". In Rath, Matthias, Friedrich Krotz and Matthias Karmasin (eds.), *Maschinenethik: Normative Grenzen Autonomer Systeme* (pp. 193-208). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21083-0_12
Westlake, Stian (ed.), 2014, *Our Work Here Is Done: Visions of a Robot Economy*. London: http://www.nesta.org.uk. http://www.nesta.org.uk
Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, . . . Jason Schultz, 2018. "AI Now Report 2018". https://ainowinstitute.org/AI_Now_2018_Report.html

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave, 2019, "Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research". 1-59. https://www.adalovelaceinstitute.org/nuffield-foundation-publishes-roadmap-for-ai-ethics-research/

Winfield, Alan, Katina Michael, Jeremy Pitt, and Vanessa Evers (eds.), 2019, *Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems* (*Proceedings of the IEEE,* Vol. 107/3). http://proceedingsoftheieee.ieee.org/upcoming-issues/machine-ethics-the-design-and-governance-of-ethical-ai-and-autonomous-systems/

Woollard, Fiona, and Frances Howard-Snyder, 2016, "Doing Vs. Allowing Harm". *Stanford Encyclopedia of Philosophy*.

Woolley, Sam, and Phil Howard (eds.), 2017, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford: Oxford University Press.

Yampolskiy, Roman V (ed.), 2018, *Artificial Intelligence Safety and Security*. London: Chapman and Hall/CRC. https://doi.org/10.1201/9781351251389

Yeung, Karen, and Martin Lodge (eds.), 2019, *Algorithmic Regulation*. Oxford: Oxford University Press.

Yudkowsky, Eliezer, 2008, "Artificial Intelligence as a Positive and Negative Factor in Global Risk ". In Bostrom, Nick and Milan M Ćirković (eds.), *Global Catastrophic Risks* (pp. 308–345). New York: Oxford University Press. https://intelligence.org/files/AIPosNegFactor.pdf

Zayed, Yago, and Philip Loft, 2019, "Agriculture: Historical Statistics". *House of Commons Briefing Paper, 3339* (25 June 2019), 1-19. https://researchbriefings.files.parliament.uk/documents/SN03339/SN03339.pdf

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan, forthcoming, "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?". *Philosophy & Technology* (First Online: 05 September 2018). https://doi.org/10.1007/s13347-018-0330-6

Zuboff, Shoshana, 2019, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

7 Academic Tools

[Auto-inserted by SEP staff]

8 Other Internet Resources

Research organizations:

Turing Institute (UK)
https://www.turing.ac.uk/media/news/alan-turing-institute-data-ethics-group/
AI Now (at NYU)
https://ainowinstitute.org/
Leverhulme Centre for the Future of Intelligence
http://lcfi.ac.uk/
Future of Humanity Institute
https://www.fhi.ox.ac.uk/
Future of Life Institute
https://futureoflife.org/
Stanford Center for Internet and Society
http://cyberlaw.stanford.edu/
Berkman Klein Center
https://cyber.harvard.edu
Digital Ethics Lab (Oxford)
http://digitalethicslab.oii.ox.ac.uk
Open Roboethics Institute
http://www.openroboethics.org/

Conferences:

Philosophy & Theory of AI
https://www.pt-ai.org/

Ethics and AI 2017
https://philevents.org/event/show/35634
FAT 2018
https://www.fatconference.org
AIES
http://www.aies-conference.com/
We Robot 2018
https://conferences.law.stanford.edu/werobot/
Robophilosophy
http://conferences.au.dk/robo-philosophy/

Policy Documents:
EUrobotics TG 'robot ethics' collection of policy documents
http://www.pt-ai.org/TG-ELS/policy
Bibliography:
PhilPapers section 'Robot Ethics'
https://philpapers.org/browse/robot-ethics
PhilPapers section 'Ethics of Artificial Intelligence'
https://philpapers.org/browse/ethics-of-artificial-intelligence

9 Related Entries

entry1 | entry2 | entry3

https://plato.stanford.edu/entries/ethics-manipulation/

https://plato.stanford.edu/entries/ethics-computer/

https://plato.stanford.edu/entries/ethics-social-networking/

https://en.wikipedia.org/wiki/Algorithmic_bias

…

<Vincent C. Müller>

<v.c.muller@tue.nl>

03.10.2019

10 Acknowledgements

Early drafts of this article were discussed with colleagues at the IDEA Centre of the University of Leeds, some friends, and my PhD students Michael Cannon, Zach Gudmunsen and Gabriela Arrieagada-Bruneau. Later drafts were made publicly available on the Internet and publicised via Twitter and e-mail to all (then) cited authors that I could locate. These later drafts were presented to audiences at the INBOTS Project Meeting (Reykjavik 2019), the Computer Science Department Colloquium (Leeds 2019), the European Robotics Forum (Bucharest 2019), the AI Lunch and the Philosophy & Ethics group (Eindhoven 2019) – many thanks for their comments.

I am grateful to the SEP reviewer, Tom Powers, for very detailed comments on the near-final draft. I am grateful for detailed comments by John Danaher, Martin Gibert, Elizabeth O'Neill, Sven Nyholm, Etienne B. Roesch, Emma Ruttkamp-Bloem, Steve Taylor, and Alan Winfield. I am grateful for further useful comments by Colin Allen, Susan Anderson, Christof Wolf-Brenner, Rafael Capurro, Mark Coeckelbergh, Yazmin Morlet Corti, Erez Firt, Vasilis Galanos, Anne Gerdes, Olle Häggström, Geoff Keeling, Karabo Maiyane, Brent Mittelstedt, Britt Östlund, Steve Petersen, Brian Pickering, Zoë Porter, Amanda Sharkey, Melissa Terras, Stuart Russell, Jan F Veneman, Jeffrey White, and Xinyi Wu.

Parts of the work on this article have been supported by the European Commission under the INBOTS project (H2020 grant no. 780073).