frontiers in Robotics and AI

REVIEW
published: 15 January 2018
doi: 10.3389/frobt.2017.00075

A Review of Future and Ethical Perspectives of Robotics and AI

*Jim Torresen**

Robotics and Intelligent Systems Group, Department of Informatics, University of Oslo, Oslo, Norway

In recent years, there has been increased attention on the possible impact of future robotics and AI systems. Prominent thinkers have publicly warned about the risk of a dystopian future when the complexity of these systems progresses further. These warnings stand in contrast to the current state-of-the-art of the robotics and AI technology. This article reviews work considering both the future potential of robotics and AI systems, and ethical considerations that need to be taken in order to avoid a dystopian future. References to recent initiatives to outline ethical guidelines for both the design of systems and how they should operate are included.

Keywords: review, ethics, technology risks, machine ethics, future perspectives

INTRODUCTION

Authors and movie makers have, since the early invention of technology, been actively predicting how the future would look with the appearance of more advanced technology. One of the first—later regarded as the father of science fiction—is the French author Jules Gabriel Verne (1828–1905). He published novels about journeys under water, around the world (in 80 days), from the earth to the moon and to the center of earth. The amazing thing is that within 100 years after publishing these ideas, all—except the latter—were made possible by the progression of technology. Although it may have happened independently of Verne, engineers were certainly inspired by his books (Unwin, 2005). In contrast to this mostly positive view of technological progress, many have questioned the negative impact that may lie ahead. One of the first science fiction feature films was Fritz Lang's 1927 German production, Metropolis. The movie's setting is a futuristic urban dystopian society with machines. Later, more than 180 similar dystopian films have followed,[1] including *The Terminator, RoboCop, The Matrix*, and *A.I.* Whether or not these are motivating or discouraging for today's researchers in robotics and AI is hard to say but at least they have put the ethical aspects of technology on the agenda.

Recently, business leaders and academics have warned that current advances in AI may have major consequences to present society:

- "*Humans, limited by slow biological evolution, couldn't compete and would be superseded by A.I.*"—Stephen Hawking in BBC interview[2] 2014.
- *AI is our "biggest existential threat,"* Elon Musk at Massachusetts Institute of Technology during an interview[3] at the AeroAstro Centennial Symposium (2014).
- *"I am in the camp that is concerned about super intelligence."* Bill Gates[4] (2015) wrote in an Ask Me Anything interview[5] on the Reddit networking site.

[1] https://en.wikipedia.org/wiki/List_of_dystopian_films.
[2] http://www.bbc.com/news/technology-30290540.
[3] https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat.
[4] http://www.bbc.com/news/31047780.
[5] https://www.reddit.com/r/IAmA/comments/2tzjp7/hi_reddit_im_bill_gates_and_im_back_for_my_third/.

OPEN ACCESS

Edited by:
Alan Frank Thomas Winfield, University of the West of England, United Kingdom

Reviewed by:
Markus Christen, University of Zurich, Switzerland
Blay Whitby, University of Sussex, United Kingdom

***Correspondence:**
Jim Torresen
jimtoer@ifi.uio.no

Specialty section:
This article was submitted to Evolutionary Robotics, a section of the journal Frontiers in Robotics and AI

Received: *05 April 2017*
Accepted: *20 December 2017*
Published: *15 January 2018*

Citation:
Torresen J (2018) A Review of Future and Ethical Perspectives of Robotics and AI. Front. Robot. AI 4:75. doi: 10.3389/frobt.2017.00075

These comments have initiated a public awareness of the potential future impact of AI technology on society and that this impact should be considered by designers of such technology. That is, what authors and movie directors propose about the future has probably less impact than when leading academics and business people raise questions about future technology. These public warnings echo publications like Nick Bostrom's (2014) book *Superintelligence: Paths, Dangers, Strategies*, where "superintelligence" is explained as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." The public concern that AI could make humanity irrelevant stands in contrast to the many researchers in the field being mostly concerned with how to design AI systems. Both sides could do well to learn from each other (Müller, 2016a,b). Thus, this article reviews and discusses published work on possibilities and prospects for AI technology and how we might take necessary measures to reduce the risk of negative impacts. This is a broad area to cover in a single article; opinions and publications on this topic come from people of many domains. Thus, this article is mostly limited to refer to work relevant for developers of robots and AI.

THE FUTURE POTENTIAL OF ROBOTICS AND AI

Many reports predict a huge increase in the number of robots in the future (e.g., MAR, 2015; IFR, 2016; SAE, 2016). In the near future, many of these will be industrial robots. However, robots and autonomous systems are gradually expected to have widespread exploitation in society in the future including self-driving vehicles and service robots at work and at home. The hard question to answer is how quickly we will see a transformation.

The technologies that surround us take many shapes and have different levels of developmental progress and impact on our lives. A coarse categorization could be the following:

- Industrial robots: these have existed for many years and have made a huge impact within manufacturing. They are mostly preprogrammed by a human instructor and consist of a robot arm with a number of degrees of freedom (Nof, 1999).
- Service robots: a robot which operates semi- or fully autonomously to perform useful tasks for humans or equipment but excluding industrial automation applications (IFR, 2017). They are currently applied in selected settings such as internal transportation in hospital, lawn mowing and vacuum cleaning.
- Artificial intelligence: software that makes technology able to adapt through learning with the target of making systems able to sense, reason, and act in the best possible way (Tørresen, 2013). There has, in recent years, been a large increase in the deployment of artificial intelligence in a number of business domains including for customer service and decision support.

The technological transition from industrial robots to service robots represents an evolution into more personalized systems with an increasing degree of autonomy. This implies flexible robots that are able to perform tasks in an unconstrained, human-centered environment (Haidegger et al., 2013). While the impact of industrial robots has been present for a number of years, the impact of service robots in workplaces and at home is still to be seen and assessed. Progress in artificial intelligence research will have a major impact on how quickly we see intelligent and autonomous service robots. Some factors that could make a contribution to this technological progress are included in Section "When and Where Will the Big Breakthrough Come?" and followed by opinions on robot designs in Section "How Similar to Humans Should Robots Become?" The possible effects of the coming technological transitions on humans and society and how to best design future intelligent systems are discussed in Section "Ethical Challenges and Countermeasures of Developing Advanced Artificial Intelligence and Robots."

When and Where Will the Big Breakthrough Come?

It is difficult to predict where and when a breakthrough will come in technology. Often it happens randomly and not linked to major initiatives and projects. Something that looks uninteresting or insignificant, can prove to be significant. Some may remember trying the first graphical web browsers that became available, such as Mosaic in 1993 (developed at the National Center for Supercomputing Applications (NCSA) at the University of Illinois Urbana-Champaign in the USA). These were slow, and it was not then obvious that the web and the Internet were something that could become as large and comprehensive as it is today. However, Internet and access to it gradually became faster and browsers also became more user friendly. So the reason why it has become so popular is probably because it is easy to use, provides quick access to information from around the world and enables free communication with anyone connected. The underlying foundation for Internet is a scalable technology being able to allow for ever-increasing traffic. For AI, the lack of technology that can handle more complex conditions has been a bottleneck (Folsom-Kovarik et al., 2016).

As the complexity of our problems increases, it will become more and more difficult to automatically create a system to handle it. Divide-and-conquer helps only to a limited extent. It remains to crack the code of how development and scaling occurs in nature (Mitchell, 2009). This applies both to the development of individual agents and the interaction between several agents. We have a lot of computing power available today, but as long as we do not know how programs should be designed, this power is limited in its contribution to effective solutions. Many laws of physics for natural phenomena have been discovered, but we have yet to really understand how complexity arises in nature. Advances in research in this area are likely to have a major impact on AI. Recent progress in training artificial neural networks with many layers (deep learning) is one example of how we can move forward in the right direction (Goodfellow et al., 2016).

In addition to computational intelligence, robots also need mechanical bodies. Their body parts are currently static after being manufactured and put in operation. However, the introduction of 3D-printing combined with rapid prototyping opens up the possibility of in-the-field mechanical reconfiguration and adaptation (Lipson and Kurman, 2012).

There are two groups of researchers that contribute to advances in AI. One group is concerned with studying biological or medical phenomena and trying to create models that best mimic them. In

this way, they try to demonstrate that the biological mechanisms can be *simulated* in computers. This is useful, notably for developing more effective medicines and treatments for disease and disability. Many researchers in medicine collaborate with computer scientists on this type of research. One example is that the understanding of the ear's behavior has contributed to the development of cochlear implants that give the deaf the sense of sounds and the ability to almost hear normally (Torresen et al., 2016).

The second group of researchers focuses more on industrial problem solving and making engineering systems sound. Here, it is interesting to see whether biology can provide inspiration for more effective methods than those already adopted. Normally, this group of scientists works at a higher abstraction level than the former group, who try to determine how to best model mechanisms in biology, but both have mutual use of each other's results. An example is the invention of the airplane that first became possible when the principle of air pressure and wing shape was understood by the Wright brothers through wind tunnel studies. Initial experiments with flexible wings similar to birds were unsuccessful, and it was necessary to have a level of abstraction over biology to create robust and functional airplanes.

Given the many recent warnings about AI, Müller and Bostrom (2016) collected opinions from researchers in the field, including highly cited experts, to get their view on the future. 170 responses out of 549 invitations were collected. The median estimate of respondents was that there is a one in two chance that *high-level machine intelligence* (defined as "a machine that can carry out most human professions at least as well as a typical human") will be developed around 2040–2050, rising to a 9 in 10 chance by 2075. These experts expect that systems will move on to *superintelligence* (defined as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest") in less than 30 years thereafter. Further, they estimate the chance is about one in three that this development turns out to be "bad" or "extremely bad" for humanity. However, we should not take this as a guarantee since predicting about the future is hard and evaluation of predictions from experts have shown that they are often wrong in their forecasts (Tetlock, 2017).

How Similar to Humans Should Robots Become?

How similar to the biological specimen can a robot become? It depends on developments in a number of fields such as AI methods, computing power, vision systems, speech recognition, speech synthesis, human–computer interaction, mechanics and actuators or artificial muscle fibers. It is definitely an interdisciplinary challenge (Bar-Cohen and Hanson, 2009).

Given that we are able to actually create human-like robots, do we want them? Thinking of humanoid robots taking care of us when we get old would probably frighten many. There is also a hypothesis called the *uncanny valley* (MacDorman and Ishiguro, 2006). It predicts that as robots get more similar to humans, the pleasure of having them around increases only until a certain point. When they are very similar to humans, this pleasure falls abruptly. Such robots might feel like the monstrous characters from sci-fi movies, and the reluctance to interact with robots *increases*. However, it later *decreases* again when they continue to be even more similar to humans; this is explained by reduced realism inconsistency (MacDorman and Ishiguro, 2006). This decrease and increase of comfort as a robot becomes more human-like is the "uncanny valley."

Although we fear the lack of human contact that could result from being surrounded by robots, for some tasks, many would prefer machines rather than humans. In contrast to most enjoying to help others, the feeling of being a burden to others is unpleasant, and we derive a sense of dignity from handling our key needs by ourselves. Thus, if a machine can help us, we prefer it in some contexts. We see this today with the Internet. Rather than asking others about how to solve a problem, we seek advice on the Internet. We probably achieve things with machines which we otherwise would not get done. Thus, in the same way as Google is helping us today with *information* needs, robots will help us with our *physical* needs. Of course, we still need human contact and social interaction. Thus, it is important that technology can support our social needs rather than making us more isolated. Autonomous cars may be one such measure, by enabling the elderly to go out and about more independently, they would support an active social life.

Whether the robots look like humans or not is less important than how well they solve the tasks we want them to handle. However, they must be easy to communicate with and easy to train to do what we want. Apple has had great success with its innovative mobile products that are easy to use. Both *design* and *usability* will be essential for many of us when we are going to choose what types of robot helpers we want in our own home in the future.

The fact that we are developing human-like robots means that they will have human-like *behavior*, but not human *consciousness*. They will be able to perceive, reason, make decisions, and learn to adapt but will still not have human consciousness and personality. There are philosophical considerations that raise this question, but based on current AI, it seems unlikely that artificial consciousness would be achieved anytime soon. There several arguments supporting this conclusion, including that consciousness can only arise and exist in biological matter (Manzotti and Tagliasco, 2008; Edelman et al., 2011; Earl, 2014). Still, robots would, through their learning and adaptation capabilities, potentially be very good at *mimicking* human consciousness (Manzotti 2013; Reggia, 2013).

ETHICAL CHALLENGES AND COUNTERMEASURES OF DEVELOPING ADVANCED ARTIFICIAL INTELLIGENCE AND ROBOTS

Ethical perspectives of AI and robotics should be addressed in at least two ways. First, the engineers developing systems need to be aware of possible ethical challenges that should be considered including avoiding misuse and allowing for human inspection of the functionality of the algorithms and systems (Bostrom and Yudkowsky, 2014). Second, when moving toward advanced autonomous systems, the systems should themselves be able to do ethical decision making to reduce the risk of unwanted behavior (Wallach and Allen, 2009).

An increasing number of autonomous systems that are working together increases the extent of any erroneous decisions made without human involvement. Several books have been published on computer ethics (also referred to as machine ethics/morality). In the book *Moral Machines* (Wallach and Allen, 2009), a hypothetical scenario is outlined where "unethical" robotic trading systems contribute to an artificially high oil price, which leads to the automated program to control energy output switches over from oil to more polluting coal power plants to avoid increasing electricity prices. Coal-fired power plants cannot tolerate running at full production long and explodes after some time and creates massive power outage with the consequences it has for life and health. Power outages trigger terror alarms at the nearest international airport resulting in chaos both at the airport and arriving aircraft colliding etc. The conclusion is that the economic and human cost was because the automated decision systems were programmed separately. This scenario shows that it is especially important for control mechanisms *between* decision systems to interact. Such systems should have mechanisms that automatically limit behavior, and also inform operators about the conditions deemed to require human review.

In the book, it is further argued that the advantages of the new technology are, at the same time, so large that both politicians and the market would welcome them. Thus, it becomes important that *morality based decision-making* becomes a part of artificial intelligence systems. These systems must be able to evaluate the ethical implications of their possible actions. This could be on several levels, including if laws are broken or not. However, building machines incorporating all the world's religious and philosophical traditions is not so easy; ethical dilemmas occur frequently.

Most engineers would probably prefer not to develop systems that could hurt someone. Nevertheless, this can potentially be difficult to predict. We can develop a very effective autonomous driving system that reduces the number of accidents and save many lives, but, on the other hand, if the system takes lives because of certain unpredictable behaviors, it would be socially unacceptable. It is also not an option to be responsible for creating or regulatory approve a system where there is a real risk for severe adverse events. We see the effect of this in the relatively slow adoption of autonomous cars. One significant challenge is that of automating moral decisions, such as the possible conflict between protecting a car's passengers relative to surrounding pedestrians (Bonnefon et al., 2016).

Below follows first an overview of possible ethical challenges we are facing with more intelligent systems and robots in our society, followed by how countermeasures related to technology risks can be taken including with machine ethics and designer precautions, respectively.

Ethical Societal Challenges Arising with Artificial Intelligence and Robots

Our society is facing a number of potential challenges from future highly intelligent systems regarding jobs and technology risks:

- Future jobs: *People may become unemployed because of automation.* This has been a fear for decades, but experience shows that the introduction of information technology and automation creates far more jobs than those which are lost (Economist, 2016). Further, many will argue that jobs now are more interesting than the repetitive routine jobs that were common in earlier manufacturing companies. Artificial intelligence systems and robots help industry to provide more cost-efficient production especially in high cost countries. Thus, the need for outsourcing and replacing all employees can be reduced. Still, recent reports have argued that in the near future, we will see overall loss of jobs (Schwab and Samans, 2016) and (Frey and Osborne, 2016). However, other researchers mistrust these predictions (Acemoglu and Restrepo, 2016). Fewer jobs and working hours for employees could tend to benefit a small elite and not all members of our society. One proposal to meet this challenge is that of a universal basic income (Ford, 2015). Further, current social security and government services rely on the taxation of human labor—pressure on this system could have major social and political consequences. Thus, we must find mechanisms to support social security in the future, these may be similar to the "robot tax" that was recently considered but rejected by the European Parliament (Prodhan, 2017).
- Future jobs: *How much and in what way are we going to work with increased automation?* If machines do everything for us, life could, in theory, become quite dull. Normally, we expect that automating tasks will result in shorter working hours. However, what we see is that the distinction between work and leisure becomes gradually less evident, and we can do the job almost from anywhere. Mobile phones and wireless broadband gives us the opportunity to work around the clock. Requirements for being competitive with others result in many today working *more* than before although with less physical effort than in jobs of the past. Although artificial intelligence contributes to the continued development of technology and this trend, we can simultaneously hope that automated agents might take over some of our tasks and thus also provide us some leisure time.
- Technology risk: *Losing human skills due to technological excellence.* The foundation for our society for hundreds of years has been training humans to make things, function, work in and understand our increasingly complex society. However, with the introduction of robots, and information and communication technology, the need for human knowledge and skills is gradually decreased with robots making products faster and more accurately than humans. Further, we can seek knowledge and be advised by computers. This lessens our need to train and utilize our cognitive capabilities regarding memory, reasoning, decision making etc. This could have a major impact on how we interact with the world around us. It would be hard to take over if the technology fails and challenging to make sure we get the best solution if only depending on information available on the web. The latter is already today a challenge with the blurred distinction between expert knowledge and alternative sources on the web. Thus, there seems to be a need for training humans also in the future to make sure that the technology works in the most effective way and that we have competence to make our own judgments about automatic decision making.

- Technology risk: *Artificial intelligence can be used for destructive and unwanted tasks.* Although mostly remotely controlled today, artificial intelligence is expected to be much applicable for future military unmanned aircrafts (drones) in air and for robots on to the ground. It saves lives in the military forces, but can, by miscalculations, kill innocent civilians. Similarly, surveillance cameras are useful for many purposes, but many are skeptical of advanced tracking of people using artificial intelligence. It might become possible to track the movement and behavior of a person moving in a range of interconnected surveillance camera and position information from the user's smartphone. The British author George Orwell (1903–1950) published in 1949 the novel "1984," where a not-so-nice future society is described: Continuous audio and video monitoring are conducted by a dictatorial government, led by "Big Brother." Today's technology is not far away from making this possible, but few fear that it will be used as in "1984" in our democratic societies. Nevertheless, disclosures (e.g., by Edward Snowden in 2013) have shown that governments can leverage technology in the fight against crime and terror at the risk of the innocent being monitored.
- Technology risk: *Successful AI can lead to the extinction of mankind?* Almost any technology can be misused and cause severe damage if it gets into the wrong hands. As discussed in the introduction, a number of writers and filmmakers have addressed this issue through dramatic scenes where technology gets out of control. However, the development of technology has not so far led to a global catastrophe. Nuclear power plants have gotten out of control, but the largest nuclear power plant accidents at Chernobyl in Russia (1986) and Fukushima in Japan (2011) were due to human and mechanical failure, not the failure of control systems. At Chernobyl, the reactor exploded because too many control rods were removed by experimentation. In Fukushima cooling pumps failed and reactors melted as a result of the earthquake and subsequent tsunami. The lesson of these disasters must be that it is important that systems have built in mechanisms to prevent human errors and help to predict risk of mechanical failure to the extent possible.

Looking back, new technology brings many benefits, and damage is often in a different form than we first would think of. Misuse of technology is always a danger, and it is probably a far greater danger than the technology itself getting out of control. An example of this is computer software which today is very useful for us in many ways, while we are also vulnerable from those who abuse the technology to create malicious software in the form of infecting and damaging virus programs. In 1999, the Melissa virus spread through e-mails leading to the failures of the e-mail systems in several large companies such as Intel and Microsoft due to overload. There are currently a number of people sharing their concerns regarding lethal autonomous weapons systems (Lin et al., 2012; Russell et al., 2015). Others argue that such systems could be better than human soldiers in some situations, if they are programmed to never break agreed laws of war representing the legal requirements and responsibilities of a civilized nation (Arkin et al., 2009).

Programs Undertaking Ethical Decision-Making

The book *Moral Machines* which begins with the somewhat frightening scenario discussed earlier in this article, also contains a thorough review of how *artificial moral agents* can be implemented (Wallach and Allen, 2009). This includes the use of ethical expertise in program development. It proposes three approaches: formal logical and mathematical ethical reasoning, machine learning methods based on examples of ethical and unethical behavior and simulation where you see what is happening by following different ethical strategies.

A relevant example is given in the book. Imagine that you go to a bank to apply for a loan. The bank uses an AI-based system for credit evaluation based on a number of criteria. If you are rejected, the question arises about what the reason is. You may come to believe that it is due to your race or skin color rather than your financial situation. The bank can hide behind saying that the program cannot be analyzed to determine why your loan application was rejected. At the same time, they might claim that skin color and race are parameters *not* being used. A system more open for inspection can, however, show that the residential address was crucial in this case. It has given the result that the selection criteria provide effects almost as if unreasonable criteria should have been used. It is important to prevent this behavior as much as possible by simulating AI systems to detect possibly unethical actions. However, an important ethical challenge related to this is determining how to perform the simulation, e.g., by whom, to what extent, etc.

It is further argued that all software that will replace human evaluation and social function should adhere to criteria such as accountability, inspectability, robustness to manipulation, and predictability. All developers should have an inherent desire to create products that deliver the best possible user experience and user safety. It should be possible to inspect the AI system, so if it comes up with a strange or incorrect action, we can determine the cause and correct the system so that the same thing does not happen again. The ability to manipulate the system must be restricted, and the system must have a predictable behavior. The *complexity* and *generality* of an AI system influences how difficult it is to deal with the above criteria. It is obviously easier and more predictable for a robot to move in a known and limited environment than in new and unfamiliar surroundings.

Developers of intelligent and adaptive systems must, in addition to being concerned with ethical issues in how they design systems, try to give the systems themselves the ability to make ethical decisions (Dennis et al., 2015). This is referred to as *computer ethics*, where one looks at the possibility of giving the actual machines ethical guidelines. The machines should be able to make ethical decisions using ethical frameworks (Anderson and Anderson, 2011). It is argued that ethical issues are too interdisciplinary for programmers alone to explore them. That is, researchers in ethics and philosophy should also be included in the formulation of ethical "conscious" machines that are targeted at providing acceptable machine behavior. Michael and Susan Leigh Anderson have collected contributions from both philosophers and AI researchers in the book *Machine Ethics* (Anderson and Anderson, 2011). The

book discusses why and how to include an ethical dimension in machines that will act autonomously. A robot assisting an elderly person at home needs clear guidelines for what is acceptable behavior for monitoring and interaction with the user. Medically important information must be reported, but at the same time, the person must be able to maintain privacy. Maybe video surveillance is desirable for the user (by relatives or others), but it should be clear to the user when and how it happens. An autonomous robot must also be able to adapt to the user's personality to have a good dialog.

Other work focuses on the importance of providing robots with internal models to make them self-aware which will lead to enhanced safety and potentially also ethical behavior in Winfield (2014). It could also be advantageous for multiple robots to share parts of their internally modeled behavior with each other (Winfield, 2017). Self-awareness regards either knowledge about one's self—private self-awareness—or the surrounding environment—public self-awareness (Lewis et al., 2015)—and is applicable across a number of different application areas (Lewis et al., 2016). The models can be organized in a hierarchical and distributed manner (Demiris and Khadhouri, 2006). Several works apply artificial reasoning to verify whether a robotic behavior satisfies a set of predetermined ethical constraints which, to a large extent, have been defined by a symbolic representation using logic (Arkin et al., 2012; Govindarajulu and Bringsjord, 2015). However, future systems would probably combine the programmed and machine learning approach (Deng, 2015).

While most work on robot ethics is tested by simulation, there are some work that has been implemented on real robots. An early example was a robot programmed to decide on whether to keep reminding a patient to take medicine, and when to do so, or to accept the patient's decision not to take the medication (Anderson and Anderson, 2010). The robot (Nao from Aldebaran Robotics) was said to make the following compromises: "Balance three duties: ensuring that the patient receives a possible benefit from taking the medication; preventing the harm that might result from not taking the medication; and respecting the autonomy of the patient (who is assumed to be adult and competent)." The robot notifies the overseer when it gets to the point that the patient could be harmed, or could lose considerable benefit, from not taking the medication. In Winfield et al. (2014) an ethical action selection mechanism in an *e-puck* mobile robot is applied to make it sometimes choose actions that compromise the robot's own safety in order to prevent a second robot from coming to harm. This represents a contribution toward making robots that are ethical, as well as safe.

Implementing ethical behavior in robots inspired by the simulation theory of cognition has also been proposed (Vanderelst and Winfield, 2017). This is by utilizing internal simulations of a set of behavioral alternatives, which allow the robot to simulate actions and predict their consequences. Using this concept, it has been demonstrated that the humanoid Nao robot can behave according to Asimov's laws of robotics.

Ethical Guidelines for Robot Developers

Professor and science fiction writer Isaac Asimov (1920–1992) was already in 1942 foresighted to see the need for ethical rules for robot behavior. Subsequently, his three rules (Asimov, 1942) have often been referenced in the science fiction literature and among researchers who discuss robot morality:

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

It has later been argued that such simple rules are not enough to avoid robots resulting in harm (Lin et al., 2012). José Maria Galvan and Paolo Dario gave birth to Technoethics, and the term was used in a talk by Galvan at the Workshop "Humanoids, A Techno-ontological Approach" at Waseda University in 2001—organized by Paolo Dario and Atsuo Takanishi—where he spoke about the ethical dimension of technology (Veruggio, 2005). The term *roboethics* was introduced in 2002 by the Italian robot scientist Gian Marco Veruggio (Veruggio and Operto, 2008). He saw a need for development guidelines for robots contributing to making progress in the human society and help preventing abuse against humanity. Veruggio argues that ethics are needed for robot designers, manufacturers and users. We must expect that the robots of the future will be smarter and faster than the people they should obey. It raises questions about safety, ethics and economics. How do we ensure that they are not being misused by persons with malicious intent?

Is there any chance that the robots themselves, by understanding that they are superior to humans, would try to enslave us? We are still far from the worst scenarios that are described in books and movies, yet there is reason to be alert. First, robots are mechanical systems that might unintentionally hurt us. Then, with an effective sensory system, there is a danger that the collected information can be accessed by unauthorized people and be made available to others through the Internet. Today this is a problem related to intrusion on our computers, but future robots may be vulnerable to hacking as well. This would present be a challenge for robots that collect a lot of audio and video information from our homes. We would not like to be surrounded by robots unless we are sure that sensor data are staying within the robots only.

Another problem is that robots could be misused for criminal activities such as burglary. A robot in your own home could either be reprogrammed by people with criminal intent or they might have their own robots carry out the theft. So, having a home robot connected to the Internet will place great demands on security mechanisms to prevent abuse. Although we must assume that anyone who develops robots and AI for them has good intentions, it is important that the developers also have possible abuse in mind. These intelligent systems must be designed so that the robots are friendly and kind, while difficult to abuse for malicious actions in the future.

Part of the robot-ethics discussion concerns military use (see Part III, Lin et al., 2012). That is, e.g., applying robots in military activities have ethical concerns. The discussion is natural for several reasons including that military applications are an important

driving force in technology development. At the same time, military robot technology is not all negative since it may save lives by replacing human soldiers in danger zones. However, giving robotic military systems too much autonomy increases the risk of misuse including toward civilians.

In 2004 the first international symposium on roboethics was held in Sanremo, Italy. The EU has funded a research program, ETHICBOTS, where a multidisciplinary team of researchers was to identify and analyze techno-ethical challenges in the integration of human and artificial entities. The *European Robotics Research Network (Euronet)* funded the project *Euronet Roboethics Atelier* in 2005, with the goal of developing the first roadmap for roboethics (Veruggio, 2006). That is, undertaking a systematic assessment of the ethical issues surrounding robot development. The focus of this project was on human ethics for designers, manufacturers, and users of robots. Here are some examples of recommendations made by the project participants for commercial robots:

- *Safety*. There must be mechanisms (or opportunities for an operator) to control and limit a robot's autonomy.
- *Security*. There must be a password or other keys to avoid inappropriate and illegal use of a robot.
- *Traceability*. As with aircraft, robots should have a "black box" to record and document their own behavior (Winfield and Jirotka, 2017).
- *Identifiability*. Robots should have serial numbers and registration number similar to cars.
- *Privacy policy*. Software and hardware should be used to encrypt and password protect sensitive data that the robot needs to save.

The studies of ethical and social implications of robotics continue and books and articles disseminate recent findings (Lin et al., 2012). It is important to include the user in the design process and several methodologies have been proposed. Value-sensitive design is one consisting of three phases: conceptual, empirical, and technical investigations accounting for human values. The investigations are intended to be iterative, allowing the designer to modify the design continuously (Friedman et al., 2006).

The work has continued including with the publications of the Engineering and Physical Sciences Research Council (a UK government agency) Principles of Robotics in 2011 (EPSRC, 2011). They proposed regulating robots in the real world with the following rules (Boden et al., 2017; Prescott and Szollosy, 2017):

1. Robots are multiuse tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws and fundamental rights and freedoms, including privacy.
3. Robots are products. They should be designed using processes which assure their safety and security.
4. Robots are manufactured artifacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
5. The person with legal responsibility for a robot should be attributed.

Further, the British Standards Institute has published the world's first *standard on ethical guidelines* for the design of robots: BS8611, in April 2016 (BSI, 2016). It has been prepared by a committee of scientists, academics, ethicists, philosophers and users to provide guidance on potential hazards and protective measures for the design of robots and autonomous systems being used in everyday life. This was followed by the IEEE Standards Association initiative on AI and Autonomous System ethics publishing an *Ethical Aligned Design, version* 1 being a *"A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems"* (IEEE, 2016; Bryson and Winfield, 2017). It consists of eight sections, each addressing a specific topic related to AI and autonomous systems that has been discussed by a specific committee of the IEEE Global Initiative. The theme for each of the sections is as follows:

1. General principles.
2. Embedding values into autonomous intelligent systems.
3. Methodologies to guide ethical research and design.
4. Safety and beneficence of artificial general intelligence and artificial superintelligence.
5. Personal data and individual access control.
6. Reframing autonomous weapons systems.
7. Economics/humanitarian issues.

The document will be revised based on an open hearing with deadline April 2017.

Civil law rules for robotics have also been discussed within the European Community resulting in a published European Parliament resolution (EP, 2017). Furthermore, discussing principles for AI were the target for the *Asilomar* conference gathering leaders in economics, law, ethics, and philosophy for five days dedicated to beneficial AI. It resulted in 23 principles within Research issues; Ethics and Values; and Longer-term Issues, respectively (Asilomar, 2017). They are published on the web and have later been endorsed by a number of leading researchers and business people. Similarly, the Japanese Society for Artificial Intelligence has published nine Ethical Guidelines (JSAI, 2017).

All the initiatives above indicate a concern around the world for the future of AI and robotics technology and a sincere interest in having the researchers themselves contribute to the development of technology that is in every way favorable.

DISCUSSION

Technology may be viewed and felt like a wave hitting us whether we want it or not. However, many novel and smart devices have been introduced that, through lack of adoption, has resulted in rapid removal from the market. Thus, through what we buy and apply, we have a large impact on what technology that will be adopted and sustained in our society. At the same time, we have limited control over unintentional changes to our behavior by the way we adopt and use technology, e.g., smartphones and the Internet have in many ways changed the way we live our lives

and interact with others. Smartphones have also resulted in us being more physically close to technology than any other living being.

In the future, there will be an even more diverse set of technologies surrounding us including for taking care of medical examination, serving us and taking us where we want to go. However, such devices and systems would need to behave properly for us to want them close by. If a robot hits us unintentionally or works too slowly, few would accept it. Mechanical robots with the help of artificial intelligence can be designed to learn to behave in a *friendly* and *user adapted* way. However, they would need to contain a lot of sensors similar to our smartphone, and we need some assurance that this *data* will *not be misused*. There are also a number of other possible risks and side effects so the work undertaken in a number of committees around the world (referred to in the previous section) is regarded as important and valuable for developing future technology. Still, there is a large divide between current design challenges and science fiction movies' dystopian portrayal of how future technology might impact or even eradicate humanity. However, the latter probably has a positive effect on our awareness of possible vulnerability that should be addressed in a proactive way. We now see this taking place in the many initiatives to define regulations for AI and robots.

Robots for the elderly living at home is a relevant example to illustrate some of the opportunities and challenges that we are facing. While engineers would work on making intelligent and clever robots, it will be up to the politicians and governments through laws and regulation to limit unwanted changes in the society. For example, their decisions are important for deciding the staff requirements for elderly care when less physical work with elderly is needed. Decisions should build on studies seeking to find the best compromise between dignity and independence on one hand and possible loneliness on the other. At the same time, if robots assume many of our current jobs, people may in general have more free time that could be well spent with the elderly.

A robot arriving in our home can start learning about our behavior and preferences and, like a child, gradually personalize its interactions, leading us to enjoy having it around similarly to having a cat or dog. However, rather than us having to take it out for fresh air, it will take us out for both fresh air and seeing friends as we get old. The exploitation of robots within elderly care is unlikely to have a quick transition. Thus, today's elderly do not have to worry about being placed under machine care. Rather, those of us who are younger, including current developers of elderly care robots, are more likely to be confronted with these robots when we get old in the future. Thus, it is in our own interest to make them user friendly.

CONCLUSION

The article has presented some perspectives on the future of AI and robotics including reviewing ethical issues related to the development of such technology and providing gradually more complex autonomous control. Ethical considerations should be taken into account by designers of robotic and AI systems, and the autonomous systems themselves must also be aware of ethical implications of their actions. Although the gap between the dystopian future visualized in movies and the current real world may be considered large, there are reasons to be aware of possible technological risks to be able to act in a proactive way. Therefore, it is appreciable, as outlined in the article, that many leading researchers and business people are now involved in defining rules and guidelines to ensure that future technology becomes beneficial to the limit the risks of a dystopian future.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

This work is partially supported by The Research Council of Norway as a part of the Engineering Predictability with Embodied Cognition (EPEC) project, under grant agreement 240862; Multimodal Elderly Care systems (MECS) project, under grant agreement 247697. I'm thankful for important article draft comments and language corrections provided by Charles Martin. Collaboration on Intelligent Machines (COINMAC) project, under grant agreement 261645

REFERENCES

Acemoglu, D., and Restrepo, P. (2016). "The race between machine and man: implications of technology for growth, factor shares and employment," in *NBER Working Paper No. 22252*. Available at: https://www.nber.org/papers/w22252.pdf

Anderson, M., and Anderson, S. L. (2010). Robot be good. *Sci. Am.* 303, 72–77. doi:10.1038/scientificamerican1010-72

Anderson, M., and Anderson, S. L. (2011). *Machine Ethics*. New York: Cambridge University Press.

Arkin, R. C., Ulam, P., and Duncan, B. (2009). *An Ethical Governor for Constraining Lethal Action in an Autonomous System*. Technical Report GIT-GVU-09-02.

Arkin, R. C., Ulam, P., and Wagner, A. R. (2012). Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. *Proc. IEEE* 100, 571–589. doi:10.1109/JPROC.2011.2173265

Asilomar. (2017). Available at: https://futureoflife.org/ai-principles/

Asimov, I. (1942). "Runaround," in *Astounding Science Fiction*, Vol. 29, No. 1. Available at: http://www.isfdb.org/cgi-bin/pl.cgi?57563

Bar-Cohen, Y., and Hanson, D. (2009). *The Coming Robot Revolution*. New York: Springer.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of robotics: regulating robots in the real world. *Conn. Sci.* 29, 124–129. doi:10.1080/09540091.2016.1271400

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science* 352, 1573–1576.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, N., and Yudkowsky, E. (2014). "The ethics of artificial intelligence," in *The Cambridge Handbook of Artificial Intelligence*, eds F. Keith and M. William (Ramsey: Cambridge University Press), 2014.

Bryson, J., and Winfield, A. F. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer* 50, 116–119. doi:10.1109/MC.2017.154

BSI. (2016). *Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems*, Vol. BS 8611 (BSI Standards Publications),

2016. Availble at: http://shop.bsigroup.com/ProductDetail?pid=000000000030320089

Demiris, Y., and Khadhouri, B. (2006). Hierarchical attentive multiple models for execution and recognition of actions. *Rob Auton Syst* 54, 361–369. doi:10.1016/j.robot.2006.02.003

Deng, B. (2015). Machine ethics: the robot's dilemma. *Nature* 523, 20–22. doi:10.1038/523024a

Dennis, L. A., Fisher, M., and Winfield, A. F. T. (2015). Towards verifiably ethical robot behaviour. *CoRR* abs/1504.03592. Availble at: http://arxiv.org/abs/1504.03592

Earl, B. (2014). The biological function of consciousness. *Front. Psychol.* 5:697. doi:10.3389/fpsyg.2014.00697

Economist. (2016). Artificial intelligence: the impact on jobs – automation and anxiety. *Economist*. June 25th 2016. Available at: https://www.economist.com/news/special-report/21700758-will-smarter-machines-cause-mass-unemployment-automation-and-anxiety

Edelman, G. M., Gally, J. A., and Baars, B. J. (2011). Biology of consciousness. *Front. Psychol.* 2:4. doi:10.3389/fpsyg.2011.00004

EP. (2017). *European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*. Availble at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-/EP/TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0/EN

EPSRC. (2011). *Principles of Robotics, EPSRC and AHRC Robotics Joint Meeting*. Availble at: https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Folsom-Kovarik, J. T., Schatz, S., Jones, R. M., Bartlett, K., and Wray, R. E. (2016). *AI Challenge Problem: Scalable Models for Patterns of Life,* Vol. 35, No. 1. Available at: https://www.questia.com/magazine/1G1-364691878/ai-challenge-problem-scalable-models-for-patterns

Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York: Basic Books.

Frey, C. B., and Osborne, M. (2016). *Technology at Work v2.0: The Future Is Not What It Used to Be*. Oxford Martin School and Citi. Available at: http://www.oxfordmartin.ox.ac.uk/publications/view/2092

Friedman, B., Kahn, P. H. Jr., Borning, A., and Kahn, P. H. (2006). "Value sensitive design and information systems," in *Human-Computer Interaction and Management Information Systems: Foundations*, eds P. Zhang, and D. Galletta (New York: ME Sharpe), 348–372.

Goodfellow, I., Yoshua, B., and Aaron, C. (2016). *Deep Learning*. Cambridge, US: MIT Press.

Govindarajulu, N. S., and Bringsjord, S. (2015). "Ethical regulation of robots must be embedded in their operating systems," in *A Construction Manual for Robots' Ethical Systems* ed. R. Trappl (Springer), 85–99.

Haidegger, T., Barreto, M., Gonçalves, P., Habib, M. K., Veera Ragavan, S. K., Li, H. (2013). Applied ontologies and standards for service robots. *Rob. Auton. Syst.* 61, 1215–1223. doi:10.1016/j.robot.2013.05.008

IEEE. (2016). *Ethical Aligned Design*. IEEE Standards Association. Available at: http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf

IFR. (2016). *World Robotics Report, 2016*. International Federation of Robotics.

Intl. Federation of Robotics (IFR). (2017). *Service Robots*. Available at: http://www.ifr.org/service-robots/

JSAI. (2017). *Japanese Society for Artificial Intelligence Ethical Guidelines*. Available at: http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf

Lewis, P. R., Chandra, A., Funmilade, F., Glette, K., Chen, T., Bahsoon, R., et al. (2015). Arch. aspects of self-aware and self-expressive comp. syst.: from psychology to engineering. *IEEE Comput.* 48, 62–70. doi:10.1109/MC.2015.235

Lewis, P. R., Platzner, M., Rinner, B., Tørresen, J., and Yao, X. (eds) (2016). *Self-Aware Computing Systems*. Switzerland: Springer.

Lin, P., Abney, K., and Bekey, G. A. (eds) (2012). *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge, Massachusetts. London, England: The MIT Press.

Lipson, H., and Kurman, M. (2012). *Fabricated: The New World of 3D Printing*. Hoboken, US: Wiley Press.

MacDorman, K. F., and Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interact. Stud.* 7, 297–337. doi:10.1075/is.7.3.03mac

Manzotti, R. (2013). Machine consciousness: a modern approach. *Nat. Intell. INNS Mag.* 2, 7–18.

Manzotti, R., and Tagliasco, V. (2008). Artificial consciousness: a discipline between technological and theoretical obstacles. *Artif. Intell. Med.* 44, 105–117. doi:10.1016/j.artmed.2008.07.002

MAR. (2015). *Robotics 2020 Multi-Annual Roadmap for Robotics in Europe*. SPARC Robotics, euRobotics AISBL. Available at: https://eu-robotics.net/sparc/upload/Newsroom/Press/2016/files/H2020_Robotics_Multi-Annual_Roadmap_ICT-2017B.pdf

Mitchell, M. (2009). *Complexity: A Guided Tour*. New York, NY: Oxford University Press, 2009.

Müller, V. C. (2016b). "Editorial: risks of artificial intelligence," in *Risks of Artificial Intelligence*, ed. V. C. Müller (London: CRC Press – Chapman & Hall), 1–8.

Müller, V. C. (ed.) (2016a). *Risks of Artificial Intelligence*. London: Chapman & Hall – CRC Press, 292.

Müller, V. C., and Bostrom, N. (2016). "Future progress in artificial intelligence: a survey of expert opinion," in *Fundamental Issues of Artificial Intelligence*, ed. V. C. Müller (Berlin: Synthese Library, Springer), 553–571.

Nof, S. Y. (ed.) (1999). *Handbook of Industrial Robotics*, 2nd Edn. Hoboken, US: John Wiley & Sons, 1378.

Prescott, T., and Szollosy, M. (2017). Ethical principles of robotics special issue. *Conn. Sci.* 29. Part 1: Available at: http://www.tandfonline.com/toc/ccos20/29/2?nav=tocList, Part2: Available at: http://www.tandfonline.com/toc/ccos20/29/3?nav=tocList

Prodhan, G. (2017). *European Parliament Calls for Robot Law, Rejects Robot Tax*. Reuters. Available at: http://www.reuters.com/article/us-europe-robots-lawmaking-idUSKBN15V2KM

Reggia, J. A. (2013). The rise of machine consciousness: studying consciousness with computational models. *Neural Netw.* 44, 112–131. doi:10.1016/j.neunet.2013.03.011

Russell, S., Hauert, S., Altman, R., and Veloso, M. (2015). Robotics: ethics of artificial intelligence. *Nature* 521, 415–418. doi:10.1038/521415a

SAE. (2016). "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," in *SAE J3016 Standard 2016* (SAE International). Available at: http://standards.sae.org/j3016_201609/

Schwab, K., and Samans, R. (2016). "The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution," in *Global Challenge Insight Report* (World Economic Forum). Available at: http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

Tetlock, P. E. (2017). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Tørresen, J. (2013). *What is Artificial Intelligence*. Oslo: Norwegian, Universitetsforlaget (Hva-er-bokserien).

Torresen, J., Iversen, A. H., and Greisiger, R. (2016). "Data from Past Patients used to Streamline Adjustment of Levels for Cochlear Implant for New Patients," in *Proc. of 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, eds J. Yaochu and K. Stefanos (Athens: IEEE Conference Proceedings).

Unwin, T. (2005). Jules Verne: negotiating change in the nineteenth century. *Sci Fiction Stud* XXXII, 5–17. Available at: http://jv.gilead.org.il/sfs/Unwin.html

Vanderelst, D., and Winfield, A. (2017). An architecture for ethical robots inspired by the simulation theory of cognition. *Cogn. Syst. Res.* Available at: http://eprints.uwe.ac.uk/31758

Veruggio, G. (2005). "The birth of roboethics," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)* (Barcelona: Workshop on Robo-Ethics), 2005.

Veruggio, G. (2006). "The EURON roboethics roadmap," in *2006 6th IEEE-RAS International Conference on Humanoid Robots*, Vol. 2006 (Genova), 612–617.

Veruggio, G., and Operto, F. (2008). "Roboethics: social and ethical implications," in *Springer Handbook of Robotics*, eds B. Siciliano and O. Khatib (Berlin, Heidelberg: Springer), 1499–1524.

Wallach, W., and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Winfield, A. F. (2014). "Robots with internal models: a route to self-aware and hence safer robots," in *The Computer after Me: Awareness and Self-Awareness in Autonomic Systems*, 1st Edn, ed. J. Pitt (London: Imperial College Press), 237–252.

Winfield, A. F. (2017). "When robots tell each other stories: the emergence of artificial fiction," in *Narrating Complexity*, eds R. Walsh and S. Stepney (Springer). Available at: http://eprints.uwe.ac.uk/30630

Winfield, A. F., Blum, C., and Liu, W. (2014). "Towards an ethical robot: internal models, consequences and ethical action selection," in *Advances in Autonomous Robotics Systems*, eds M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish (Springer), 85–96.

Winfield, A. F., and Jirotka, M. (2017). "The case for an ethical black box," in *Towards Autonomous Robot Systems*, ed. Y. Gao (Springer), 1–12. Available at: http://eprints.uwe.ac.uk/31760

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.