Tallinn
Digital
Summit

digitalsummit.ee

Briefing note: Responsible, Safe and Secure AI

By Luukas Ilves, the Lisbon Council

theLisbonCouncil
think tank for the 21st century

This briefing note has been prepared by the Lisbon Council for Tallinn Digital Summit, October 2018.

Intro

In the coming decade, as many as 50% of work tasks will be subject to automation by the use of Artificial Intelligence (AI).[1] AI will enable scientific breakthroughs, help us conquer disease and climate change and transform every sector of the economy. The development and adoption of AI is an economic imperative and an arena of active geopolitical competition.

For all their capabilities, AI tools are still imperfect. They make errors in judgment, act unpredictably and can be tricked. As we take humans out of the loop of some decision making, we are replacing them with agents that possess neither common sense nor a conscience. These are sophisticated optimisation functions that are only as good as their programming and the data used to train them, they are prone to new kinds of errors and can be tricked.

Overreliance on AI can be dangerous, leading to accidents and a breakdown in legal liability and responsibility, while undermining human rights and due process. And AI tools will also be put to malicious use to amplify cyber attacks, manipulate human psychology and as a tool of warfare. Eminent thinkers have gone so far as to label AI an existential threat to humanity.

The good news is that a wide range of discussions and initiatives across disciplines and domains is starting to look at these questions. Governments, major companies, NGOs and researchers are all getting in on the act, proposing new norms, laws and mechanisms.

This is an urgent discussion that needs to be held while the technology is still new and being put into place. We can pre-empt and avoid the most serious threats and avoid repeating mistakes of the past. This paper looks at what steps can help ensure a responsible use of AI, concluding with a practical roadmap for policymakers, executives and researchers, merging recommendations from across a wide range of disciplines and discussions.

A multidisciplinary, generational challenge will require multi-disciplinary solutions. Legal norms and consumer and market pressures will need to function together to encourage new technologies and research. Companies will need to learn to develop AI responsibly, while policymakers will need to tolerate some risk and uncertainty. Done right, this balance can achieve can lead to an era of AI that contributes to human flourishing.

[1] James Manyika, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko and Saurabh Sanghvi, *Jobs Lost, Jobs Gained: What the Future of Work Will Mean for Jobs, Skills and Wages* (Washington: McKinsey Global Institute, 2017).

1. AI: Hyper-competent and fallible

The term "artificial intelligence" was coined in the 1950s. While a comprehensive definition is hard to pin down, the term broadly refers to machines that mimic functions we associate with human intelligence, such as learning or problem solving, or exhibit a high degree of autonomy in their action.[2]

The development of AI technology has gone through several periods of optimism followed by periods of slowdown ("AI winters"). Most implementations of AI today rely on machine learning, the use of algorithms that learn from data to make predictions and find new patterns and pathways. Machine learning mimics how biological organisms learn but can do so at the speed of computers. In the last decade, vastly increased computing capacity and specialised processors paired with exponentially larger datasets have made for major breakthroughs, and AI applications have surpassed numerous benchmarks of human intelligence. Improvements in speech and image recognition now enable machines to interact with humans and navigate the physical world, while deep learning techniques have enabled AI to match human reasoning in many domains.

AI-s now outperform humans at a growing list of tasks, finding novel strategies and moving much faster than people. AI-s can beat us at most of our games (Chess, GO and even Jeopardy), detect cancer with a lower error rate than trained experts, and engage in creative tasks like composing new music. Using AI, computer programmes are even able to alter their own programming. Yet AI systems remain fallible in several important ways:

AI is only as good as the data or environment it learns from. An AI learns from its "training data" and past experience and will incorporate biases or errors into its picture of reality. AI tools frequently end up replicating existing human biases. For instance, a 2015 study showed that Google displayed high-paying jobs advertising to men at significantly higher rates than to women, most likely not due to explicit discrimination on anyone's part, but because the algorithm had learned from past experience that men were more likely to click on such advertisements.[3] And partial data sets can limit the efficacy of AI tools. A speech recognition algorithm that has only learned from a particular accent might struggle to recognise that language with other accents. In another case, an algorithm designed to identify tanks was given training data where images of tanks were taken on cloudy days, while photos without tanks were sunny, so the algorithm instead learned to distinguish between sunny and cloudy days.[4]

KRATT

Mythology and literature have run far in advance of current debates. Old Estonian mythology tells of *kratts*. A *kratt* was formed from hay or old household implements by its master, who gave the devil three drops of blood to bring the creature to life. Cunning peasants are said to have used blackcurrant berries in lieu of blood to cheat the Devil and save their souls from going to hell.

The *kratt* did everything the master ordered it to and was mostly used for stealing and bringing various goods to its owner, though *kratts* were prone to misunderstanding instructions. They could also fly. The *kratt* needed to be kept constantly working, lest it turn dangerous to its owner. Once a *kratt* became unnecessary, its master would give it impossible tasks, which would ultimately lead it to catch fire and burn to pieces. For more, visit the Estonian Culture Centre Website at http://www.rahvakultuur.ee/Welcome_to_the_website_of_Estonian_108.

[2] For instance, the Gartner IT Glossary offers the following definition of Artificial Intelligence: technology that appears to emulate human performance typically by learning, coming to its own conclusions, appearing to understand complex content, engaging in natural dialogues with people, enhancing human cognitive performance (also known as cognitive computing) or replacing people on execution of nonroutine tasks https://www.gartner.com/it-glossary/.

[3] The Google example was discovered by researchers conducting testing of Google's services, see Amit Datta, Anupam Datta, Ariel D. Procaccia and Yair Zick, "Influence in Classification via Cooperative Game Theory," *International Joint Conferences on Artificial Intelligence*, pp. 511-517, 2015. Roman Yampolskiy and MS Spellchecker "Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures," *arXiv preprint arXiv:1610.07997*, 2016 has a long list of examples of AI failures.

[4] A wide range of AI applications are trained on a relatively small number of freely available data sets, many of which are listed here: https://gengo.ai/datasets/the-50-best-free-datasets-for-machine-learning/.

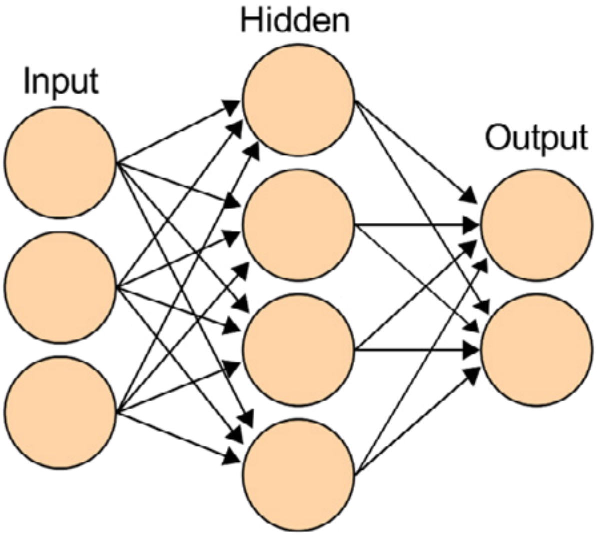
Tank (L), No tank (R)

AI machines lack what we would call "common sense." Machine and deep learning optimise a mathematical equation, without a sense of perspective or an innate secondary order sense of when things "aren't right." When creating AI-based programmes, safeguards need to be explicitly designed in. When constraints are left un-specified, unexpected results can ensue. For instance, a conversational chatbot developed by Microsoft (named "Tay") learned from its interactions with human online conversation partners to use racist and sexually abusive language, leading Microsoft to take to Tay offline within 16 hours of its launch.[5]

Even when they work well, sophisticated AI-s have become harder to understand, not just for the end users, but even for the people who designed them in the first place.[6] Deep learning, a sophisticated form of machine learning, uses neural networks and evolutionary algorithms, which are essentially "AI-s being built by AI-s", that can quickly resemble a tangled mess of connections that are nearly impossible for analysts to disassemble and fully understand.

Neural networks can identify patterns and trends among data that would be too difficult or time-consuming to deduce through human research. In its defeat of Go champion Lee Sedol, Google's AlphaGo AI used moves that "human players would never think about doing."[7]

The black box nature of advanced AI can create difficulties in processes where we value predictability or understanding the underlying reasons for an action. If one cannot interrogate an AI as to why it made a certain determination, it will be difficult for us to build AI into critical processes. Research into self-explaining AI, which uses various techniques to present the factors that affected its analysis to humans, promises to overcome some of this gap.

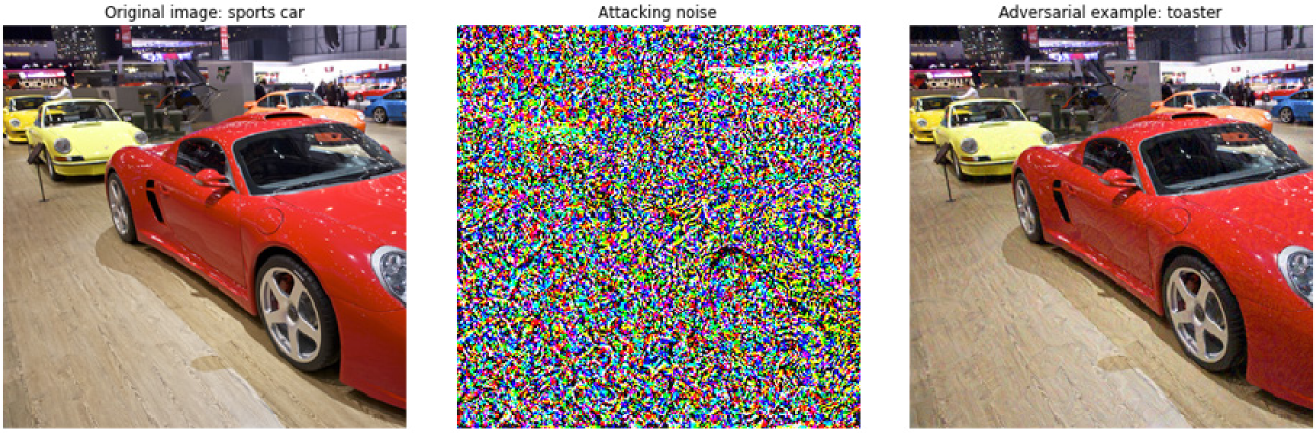


A neural network.
Image by Colin Burnett

[5] See Hope Reese, "Why Microsoft's 'Tay''AI Bot Went Wrong," *Tech Republic*, 24 March 2016. https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

[6] Martin Wekler, „Building AI Systems that Work is Still Hard," *Tech Crunch*, 01 August 2018 https://techcrunch.com/2018/01/01/building-ai-systems-that-work-is-still-hard/

[7] Joon Ian Wong and Nikhil Sonnad, "Google's AI Won the Game Go by Defying Millennia of Basic Human Instinct," *QZ.com*, 25 March 2016.



Example of noise introduced into an adversarial attack leading an AI to misclassify a sports car as a toaster, while leaving the image unaltered to human eyes.

Image credit: Jesus Rodriguez

AI is also vulnerable in new ways and can be fooled in ways that humans can't. Algorithms are susceptible to "adversarial attacks," illusions designed to fool machine-learning algorithms into doing things like mistaking a picture of a car for a toaster. They can be images, sounds or paragraphs of text — and recent research is showing that such attacks are much easier to pull off than previously understood. While a car-toaster mix-up may seem low stakes, an adversarial example could attack the AI system that controls a self-driving car, for instance, causing it to mistake a stop sign for a speed limit. Such adversarial attacks have already been used to beat other kinds of algorithms, like spam filters.[8]

2. AI among humans

From the early days of computing in the 1940s, the sophistication of automated systems has brought about unintended consequences, from the amusing to the terrifying. We expect AI systems to do things better than people, but they do not always act as we expect them to. The danger arises when we do not recognise an AI system's weakness.

An incident from the Cold War illustrates the potential cost of over-relying on automated systems. On 26 September 1983, the Soviet nuclear early-warning system calculated that the United States had launched multiple Minuteman intercontinental ballistic missiles toward the Soviet Union. This would have called for an immediate retaliatory launch, but lieutenant colonel Stanislav Petrov sensed that something was "off" — why would the US launch a first strike with only four or five missiles? He correctly identified the alert as a false alarm, and quite possibly saved us all from nuclear annihilation.[9]

The behaviour of autonomous systems using AI is a complex interaction of initial programming and algorithms, training data, interaction with the physical world and human beings. The factors that can lead to such accidents are often unpredictable.

Liability and the Law

New or updated legal regimes may help bring some clarity to the question of AI liability:

The European Commission is assessing whether there are gaps in national and EU liability and safety frameworks and is considering changes to the EU Product Liability directive. For more, read European Commission, *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe*.

Estonia is considering the introduction of a law on robotics, a so called "*Kratt*-law", that would give limited legal agency and representation to robot-agents, while clarifying the liability owners, operators and manufacturers would have for the actions of robot agents. A comparison is made to animals, whose owners can be held liable for the actions of their pets. See Karmen Turk and Maarja Pild, *Analüüs SAE Taseme 4 ja 5 Sõidukite Kasutusele Võtmiseks: Kitsas ja Lai Vaade* (Tallinn: Triniti, 2017).

[8] Some of these risks may not be mitigatable at all, an unavoidable consequence of the complexity of machine learning systems, as noted by Nicolas Papernot, Patrick McDaniel, Arunesh Sinha and Michael Wellman in "Towards the Science of Security and Privacy in Machine Learning," *arXiv preprint arXiv:1611.03814*, 2016.

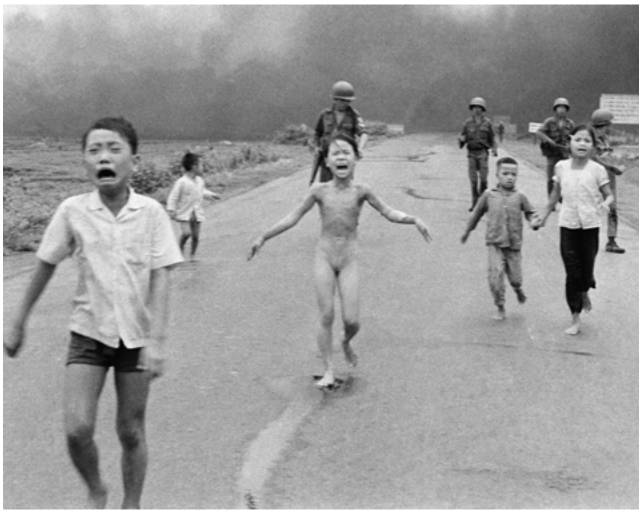
[9] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: WW Norton & Company, 2018).

3. AI in the physical world

As AI-enabled autonomous systems become more capable and replace humans in an increasing range of areas, we should expect overall safety to increase. AI-s can react faster and more accurately, while avoiding many of the sources of human error that lead to crashes and accidents of all kind, from automobiles to stock markets. Yet this development is unlikely to be entirely straightforward. Autonomous agents will have accidents, and in new and unforeseen ways.

As robots become competent in day-to-day situations, we face the problem of people coming to blindly rely on systems and being unprepared for their occasional failure. Aircraft autopilots can — in ideal circumstances — automate the entire process of flying, including take-off and landing. Yet automated aircraft systems perform far less well in extreme situations. The crash of Air France Flight 447 in 2009 (killing all people on board) is now attributed neither to malfunctioning AI nor pilot error, but an interaction of the two.[10] And there is now concern that pilot skill is atrophying, as pilots lose the habit of regularly piloting their aircraft.[11]

Today, we use a combination of regulation, liability rules and reputational pressures to provide incentives to ensure products, services and human behaviour are safe. Yet responsibility is not always straightforward to assign. Complex AI systems create new forms of the "many-hands" problem, whereby agency and responsibility for an outcome are distributed between a wide range of participants. In the case of Microsoft's racist chatbot Tay, who is responsible? The programmers? The users who interacted with Tay using racist language? Or neither?



Iconic war photo or child pornography? Facebook's algorithm flagged posts with this picture as illegal material, public outcry forced Facebook to let the offending posts back up.

Photo by Nick Ut, AP, 8.6.1972

4. Does AI know right from wrong?

AI's intrusion into the moral and ethical realm of human choices will be at least as disruptive as the changes it brings to the physical world. Algorithms and AI are already being used to make decisions with a direct impact on other people's lives and livelihood: granting parole, hiring, giving financial credit or security clearances.

At its best, AI helps us make better decisions, reducing human biases and cutting easy cases out of the workflow, allowing humans to focus on truly difficult or complex cases. But AI can also amplify existing biases — parole granting AIs have been shown to discriminate against African-Americans, and facial recognition software has a higher false-positive rate for many non-white faces, which has a potential to lead to harassment, or unfair

[10] Madeleine Claire Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," *We Robot 2016* (2016); and Matthew Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies," *Harvard Journal of Law and Technology* 29 (2015): 353 delve into these issues in greater depth. Ms Elish coins the term of "moral crumple zones" for situations where responsibility cannot properly be attributed. Mr Scherer notes that the law around these questions is very underdeveloped, with almost no AI specific laws, though robotics law has developed rules for certain circumstances.

[11] See for example Donna Mahoney, "Flying on Autopilot Improves Airlines Safety but Can Lead to Errors," *Business Insurance*, 28 February 2016. https://www.businessinsurance.com/article/00010101/NEWS06/302289983/Flying-on-autopilot-improves-airline-safety-but-can-lead-to-errors

incarceration. These are often a function of imperfect training data or misweighted algorithms, but the error can only be discovered by looking at the aggregate data on the results of automated processing.

AI can also be put to use in ways that endanger human rights such as privacy or free expression. For instance, China's social credit systems uses widespread surveillance, facial recognition and analysis of large datasets to identify and reward or punish citizens for behaviour that the government approves of or deems to depart from the ideal. And AI can be used to de-anonymise individuals in large data sets, for instance to reveal sensitive health data.

Even when AI is used with the best of intentions, outsourcing moral decision making can have unintended consequences. Social media platforms face significant pressure to remove posts containing hate speech, terrorist material, child pornography or violating copyright protections. To police billions of posts, they have turned to algorithms to identify material for takedown. Yet these machines often misfire and flag perfectly legal expression for removal. And the same technology can be used by illiberal regimes to find — and eliminate - dissent.

5. Safeguards against AI fallibility

Ethics and human rights considerations are not merely a "nice-to-have" feature. Companies can face significant legal exposure and reputational risk when they discriminate, and governments are entrusted with protecting rights and ensuring due process. Indeed, the difficulty in assigning responsibility, whether for protecting rights or civil liability, points to a broader difficulty in assuring the rule of law.

As AI becomes embedded in decision-making processes, certain types of safeguards and principles are being widely discussed as means for ensuring responsible adoption and use of AI:[12]

- **Transparency** on where algorithmic processing or AI-enabled decision-making are being used to empower those who interact with it and is the basis for good oversight.

- **Access to data** can help public authorities and watchdogs exercise oversight. All sorts of data can be helpful, including the training data used to teach an AI and outcome data about how an AI is performing in the real world. Sometimes, only testing by third parties can reveal issues.

- **Explainability** — providing explanations as to how an AI works, and making this information comprehensible to lawyers, ethicists and

AI human rights in law

AI ethics principles and rules on keeping humans in the loop are beginning to appear in legislation:

The General Data Protection Regulation (GDPR) of the European Union, which entered into force on 25 May 2018, sets out strict criteria for the processing, storing and exchange of personal data. Because of certain provisions related to algorithmic processing, the GDPR has been called the first "AI law" - the regulation enshrines the right of individuals not to be subject to decisions made on the basis of automated profiling alone — a human being must be in the loop and take ultimate responsibility for the conclusions reached. Additionally, the GDPR gives the right to challenge decisions made by automated decision making and receive an explanation for why a certain decision was made.[13]

Recently, Microsoft proposed that government step in to regulate the use of facial recognition technology, given the impact it can have on the basic enjoyment of privacy and civil liberties. Microsoft asked lawmakers to consider a number of questions around the use of facial recognition by law enforcement and in public spaces, as well as what rights individuals should have.[14] The company has argued that certain norms should not be a matter of competition between companies and are best set by governments.

[12] Much of this work originated from a research community around questions of Fairness, Accountability, Transparency and Explainability (FATE), which now meets for several annual conferences. For a reading list, see Erini Malliaraki, "Toward Ethical Transparent and Fair AI/ML: A Critical Reading List," *Blog,* 2018. https://medium.com/@eirinimalliaraki/toward-ethical-transparent-and-fair-ai-ml-a-critical-reading-list-d950e70a70ea.

[13] For an excellent overview of European Union rules on personal-data protection, visit https://ec.europa.eu/info/law/law-topic/data-protection_en.

[14] Microsoft President Brad Smith explains the rationale for the proposal in a blog post "Facial Recognition Technology: The Need for Public Regulation and Corporate Responsibility," *Microsoft Blog*, 3 July 2018.

users.[15] While AI may sometimes function as a black box, work is also being done on AI-s that can themselves explain their decisions.[16] A separate but related goal is understandability: measures to ensure humans use AI in a well-informed manner. Measures can include well-designed user interfaces and warnings about possible risks and interactions.

- **The right to redress** and human review when a decision is made based on AI, including the ability to correct input data.
- Beginning the deployment of an AI-s in a **sandbox** environment — with stringent additional safety safeguards or limitations in the testing area. A related measure is to only use AI in the types of environments it has been tested and trained for.
- Conducting an AI **impact assessment**, similar to the data protection impact assessment required from some data processors under the GDPR.
- Designers of autonomous systems should strive to make their behaviour **consistent and predictable**.

However, these measures are not free. They entail significant effort and cost on the part of both companies developing AI, regulators and users. While these measures all make up part of a useful toolbox, we need to carefully consider when to require them by law.

Ongoing testing of self-driving cars gives us an idea of what some of these measures look like in practice.[17] Self-driving cars were initially tested in closed environments, then on general roads in certain conditions (e.g. on certain roads, with human drivers present, at certain speeds). A five-tiered system is used to describe different levels of autonomy.[18] Self-driving cars are subject to stringent accident reporting and transparency requirements.[19] Courts will need to figure out when a driver is at fault for an accident, and when and how to assign responsibility to others. Over time, new rules will settle into place regarding the use of self-driving cars, including technical standards and performance requirements, interactions between human and machine, and liability.

6. Hacking with AI, Hacking AI

For much of the history of cyber conflict, the offense has held the advantage. With the sheer volume of cyber attacks in the billions every day, automation has proven a key technique for the defence to keep up. Machine learning's pattern recognition capabilities are ideally suited to learning what the regular functioning of an IT system looks like and identifying anoma-

[15] For a good explainer on explainability, see https://www.sentient.ai/blog/understanding-black-box-artificial-intelligence/.

[16] DARPA (the Defense Advanced Research Projects Agency of the United States), for instance, is sponsoring work on self-explaining AI. See https://www.darpa.mil/program/explainable-artificial-intelligence.

[17] See KPMG International, *Autonomous Vehicles Readiness Index* (Geneva: KPMG International, 2018).

[18] Developed by the Society of Automotive Engineers, known as SAE Levels 1-5. See http://www.sae.org/misc/pdfs/automated_driving.pdf.

[19] Though not all jurisdictions have applied all of these requirements. For instance, Arizona, site of a recent fatal accident, became an attractive destination for piloting due to possessing laxer testing requirements than other U.S. states like California.

lies that hint at hacking or compromise. AI can also help automate threat intelligence and even attribute attacks.[20] Every cyber security vendor now prominently advertises their machine learning and AI capabilities.

The next step lies in automating the full process of cyber defence — and offence. In 2018, IBM demonstrated an AI-powered malware toolkit that could bypass traditional detections and make the types of highly sophisticated attacks.[21] So far, we have not seen widespread use of AI by cyber attackers — perhaps because the most capable AI experts are currently working for the likes of Microsoft and Google.[22]

In the longer run, offensive cyber tools will proliferate. Many AI tools are available online as open source tools and can be paired with inexpensive malware kits to make sophisticated automated cyber attack tools.[23] And the most advanced state-developed AI cyber tools will almost inevitably spread, much as happened in the case of Stuxnet, an attack originally targeted against Iran's nuclear programme whose code has shown up in cyber attacks across the world. An optimistic view suggests that AI cyber attack tools could actually help the defence, by allowing for the stress-testing of systems against possible attacks before they go live.

The widespread use of AI itself creates a new attractive set of targets. As automation progresses, the impact of a successful cyber attack becomes increasingly profound. An attack on a self-driving car could be used to commit murder — or simply to enable a more convenient form of car theft, with the stolen car driving itself into the garage of the thief. Worryingly, as we put increasing trust in AI systems, we may not even realise that they have been hacked.

Some of the most damaging cyber attacks will be against AI systems themselves, executing adversarial attacks or attempting to alter algorithms. An attack against training data can be used to alter how an AI functions. Hackers can also use techniques to reveal information about the underlying training set or to determine whether a piece of data was part of the training data, problematic when an AI has learned from personal or sensitive data.[24]

[20] For some explanation of these capabilities, see Anna Buczak and Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys and Tutorials* 18, no. 2 (2016): 1153-1176; Eric Nunes, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart and Paulo Shakarian, "Darknet and Deepnet Mining for Proactive Cybersecurity Threat Intelligence," *arXiv preprint arXiv:1607.08583*, 2016 and Yuchi Tian, Kexin Pei, Suman Jana and Baishakhi Ray, "Deeptest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars," *Proceedings of the 40th International Conference on Software Engineering*, pp. 303-314, 2018.

[21] Kevin Townsend, "IBM Describes AI-powered Malware That Can Hide Inside Benign Applications." Security Week, 13 August 2018.

[22] This is a commonly held view expressed to the author by a number of experts in the field.

[23] Google now makes it possible to automate building your own AI system. See Will Knight, "Google's Self-Training AI Turns Coders into Machine-Learning Masters," *Technology Review*, 17 January 2018.

[24] Many of these adversarial attacks are inherently hard to defend against. Some of the first such adversarial attacks we are seeing use audio as a vector, e.g. to hack Amazon's Alexa or Siri. See Nicholas Carlini and David Wagner, "Adversarial Examples are Not Easily Detected: Bypassing Ten Detection Methods," *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017. For a full list of AI potential vulnerabilities, see Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu and Victor CM Leung, "A Survey on Security Threats and Defensive Techniques of Machine Learning: a Data Driven View," *IEEE access* 6 (2018): 12103-12117. Importantly, these attacks are not just carried out under controlled conditions; Mr Liu's team reliably fooled Google's Cloud Vision API, a machine learning algorithm used in the real world today.

Hacking People

What happens when we no longer know if we are talking to a person on the other end? Autonomous agents add a new spin to the anonymity the internet affords. Today, we might say "On the internet, nobody knows you're a machine."

AI can also be used to "hack people." The canonical test for machine intelligence is the Turing test — can a computer fool a person into thinking they are interacting with another person? AI has gotten pretty good at passing the Turing test.

Machine learning has taught chatbots to imitate the nuances of human conversation, not only using slang but peppering text with typos to appear more human. AI tools will become a useful tool for manipulating humans, whether to carry about scams or manipulate human sentiments and emotions.

"On the Internet, nobody knows you're a dog."

Cartoon by Peter Steiner, *The New Yorker*, 5.7.1993

Bots and AI are nowhere more disruptive than in the marketplace of ideas. Recent elections, peppered with "fake news," have given us a foretaste. Automated translation helps Russian trolls pose as locals. Twitter bots amplify the message of fringe figures. AI helps profile and target news and advertising at niche segments of the population most likely to react to a message, fragmenting the public sphere.

Even as democracies learn to inoculate themselves against the current generation of fake news, new challenges are on the horizon. The quality of machine-generated audio and video is expanding in leaps and bounds. In just the last year, tools to produce relatively convincing "deep fakes" — artificial video and audio — have become cheaply available, making it possible to create completely fake audio and video. We may already have seen the first cases of these tools being weaponised to discredit political opponents. What happens to "fake news" when we literally cannot believe our eyes and ears?

7. Robots at war

AI is no newcomer to the battlefield. There has been a gradual evolution of automation in weapons systems. Advanced weapons systems already acquire and engage targets with no human intervention, albeit not without incident. During the 2003 Iraq war, autonomous air defence missile systems were responsible for the majority of airborne friendly fire incidents.[25]

The world's major military powers all have significant research programmes in autonomous weapons. While the US has the lead in such research, other countries are catching up. And Vladimir Putin has said of AI that "Whoever becomes the leader in this sphere will become the ruler of the world."[26]

AI and digital identity

Anonymity is widely taken as a feature of the modern internet, but this is hardly inevitable. Online platforms and publishers can take steps to ensure the person behind an online avatar is human. In many countries, libel laws have incentivised news outlets to require online identification, and Twitter verifies the identities of some users.[27]

An increasing number of countries now have the technical infrastructure to either private- or government-issued electronic identities to physical identities, creating certainty about the identity of an individual behind an online avatar. Some notable examples include India's Aadhar, Sweden's BankID and Estonia's national electronic ID. These identities are in turn used in a wide range of interactions, from consumer banking and commenting on online news sites to government services to signing employment contracts.

The inability to distinguish between AI and humans has given rise to proposals for a new human right — the right to know whether you are interacting with a human being or an AI.

[25] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: WW Norton & Company, 2018).

[26] Op. cit.

[27] In the Delfi v. Estonia case, the European Court of Human Rights ruled that the Estonian media outlet Delfi could be held liable for libelous comments posted on their website, but noted in the judgment that identifying the physical persons behind those comments could allow Delfi to pass on liability to the actual authors of the comments. In response, most Estonian media outlets started requiring identification in order to post comments on news stories. See *Delfi v Estonia: Judgment (Application no. 64569/09)*, (Strasbourg: European Court of Human Rights, 2015).

Paired with developments in robotics, we can now imagine a wholly different kind of warfare, with swarms of drones and autonomous ground vehicles taking the place of boots on the ground. There are many reasons why AI in the battlefield presents an attractive prospect to many militaries. It can be used to save lives — both by keeping human soldiers out of the most dangerous parts of the battlefield but also avoiding collateral damage and inadvertently targeting civilians. A machine can be programmed to follow rules of engagement and the law of armed conflict faithfully. Furthermore, beyond a certain degree of automation, keeping humans in the loop becomes impossible — one of the first steps in any conflict is to jam the other side's communications, and human reaction time can be too slow.

Claims about the humane character of new weapons have been made before — the inventers of the machine gun believed it would save lives by sparing human soldiers bloody hand-to-hand conflict. And even if a new weapon does save lives, it may not make the world more peaceful. The perception that a weapon is more peaceful and humane may lead it to be used more frequently. If fights between robotic weapons are seen as a less aggressive form of warfare, this may also lower the threshold for engaging in armed conflict.

In reaction, a broad international discussion led by NGOs has begun on banning (some kinds of) autonomous or robot weapons. They worry both about how such weapons could change the face of conventional warfare, but also of the risk of loss of control, whether through hacking or autonomous agents gone rogue. And certain actors may wish to remove themselves entirely from the domain of weapons and AI. For instance, Google's AI Principles state that the company wil not design or deploy applications in the area of weapons and related technologies. The international community has been remarkably quick to act in this area: a UN group of experts has already been convened to produce recommendations.[28]

Outright bans of autonomous weapons are unlikely to work. AI arms control presents the same difficulties as cyber arms control. The basic technology is relatively accessible, and inherently dual-use in nature. Nuclear or chemical weapons require extensive physical infrastructure to develop and store (and can thus be inspected), while the autonomy and programming of a weapons system is a function of its code, which can be hidden and altered. And civilian robotics and AI can easily be combined with weapons, as in the case of handguns attached to drones.

Some of ideas being discussed to mitigate these risks include:

- A limitation on the use of autonomous weapons only to specific environments where they have been tested and evaluated, which would minimize the risk of unforeseen consequences and out-of-control AI.
- Specifying certain situations or decisions where humans have to be kept in the loop.
- Placing more restrictive combat rules of engagement on autonomous weapons — and announcing these rules of engagement publicly.
- A transparency mechanism facilitating the post-engagement review of how autonomous weapons behaved on the battlefield.

[28] The 2017 and 2018 UN Group of Governmental Experts on Lethal and Autonomous Weapons Systems

8. An existential risk?

For all the risks outlined above, might the rise of Artificial Intelligence herald something more ominous? Stephen Hawking, Elon Musk and Bill Gates have joined numerous academics and thinkers in warning that runaway AI could pose an existential risk for humanity as a whole.[29] Once Artificial Intelligence is able to fully match the capabilities of the human mind, it can marshal the mass and speed of machine computing power to run circles around human intelligence — billions of instances each in turning running billions of calculations per second. [30]

A super-intelligence would be able to control the networked physical world, hack its way around our security, and would be an adept psychological manipulator of humanity. Humans could find themselves at the mercy of this super-intelligence, with the machine able to out-think and out-do them every step of the way. Should the super-intelligence come to see humanity as a threat, it would likely win the war before we realized it had even begun. Or it might simply pursue goals orthogonal to human flourishing.

Experts differ greatly in their views over whether and when an artificial super-intelligence could be created, whether it would pose a risk, and what could be done about it. Most AI researchers dismiss these fears of an AI Armageddon as science fiction and unrelated to the practical research questions they face, yet even sceptics agree that the possibility cannot be dismissed out of hand.[31] Even if sensationalist, scenarios of out of control AI can serve a purpose — they help stoke the imagination and draw attention our dependence on autonomous IT systems, highlighting systemic risks from hacking or error. What the AI might not do, an out of control hacker or state adversary might.

Since Isaac Asimov postulated three laws of robotics, science fiction writers have been exploring how we might design truly autonomous systems so they do not harm humans and remain under our ultimate control.[32] Current research in AI control is asking the same question, looking in particular at how to ensure that robots understand the underlying intent of human commands and do not implement them perversely, code certain positive values and goals immutably into programs so they could not later be re-written and — when these safeguards fail — halt AI systems run amok. It may be that the most productive avenues are not around limiting the harm auton-

[29] A 2017 Open Letter by 116 AI researchers makes the same argument: https://futureoflife.org/autonomous-weapons-open-letter-2017/

[30] Most famously, futurologist Ray Kurzweil predicts the arrival of the singularity by 2050. Nick Bostrom's book *Superintelligence* (Paris: Dunod 2017) is the most comprehensive overview of scenarios leading to an existential threat from AI. For an overview of when researchers think "artificial general intelligence" will arrive, see Vincent Müller and Nick Bostrom Müller, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion" in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Berlin: Springer, 2016). Many who are sceptical of its arriving in the short term nevertheless agree the possibility cannot be discounted.

[31] The *Asilomar AI Principles* agree in principle that "there being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities" and "advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources." Future of Life Institute, Asilomar AI Principles (Asilomar: Future of Life Institute, 2017).

[32] The three laws (from Isaac Asimov, *I, Robot* (New York: Gnome, 1950)) are:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

omous agents might do but teaching them instead to maximize human empowerment and control, understanding the impact their actions have on their environment — including humans.[33]

9. An agenda for human control

The development of cybersecurity provides a cautionary tale of how not to handle a growing complex risk management problem. Poor cybersecurity has become a major drag on the world economy and fuels a cybercrime industry worth nearly one trillion euros, several thousand major breaches of personal data every year affect hundreds of millions of people and cyber attacks have become a regular part of international conflict.[34]

In spite of this impact, the world's collective record has been poor. People don't pay attention to security. Most cyber attacks could be stopped by basic best practices. But this is inconvenient and expensive, and the pressures in the ICT world are to move quickly and ask for forgiveness later. Additionally, there is a major dearth of cyber security experts. Laws, international norms and strategic logic have struggled to keep up. Only in the last few years have countries extensively adopted national cyber strategies and passed legislation.

The task of adopting AI in a manner that is safe to use is a generational challenge, on par in breadth and importance with environmental sustainability.[35] We do not have the luxury of repeating past procrastination.

The good news is that multiple communities are currently engaged in far-ranging technical and policy debates across the issues raised in this paper. From human rights NGOs and large corporations to defence ministry legal departments, they approach these questions from radically different vantage points. It is all the more remarkable, then, that many of the measures being considered in these discussions are similar and even converge, pointing at a common agenda.

We would do well to bear in mind some important first principles:[36]

- First and foremost, human beings are still in charge. We will create the framework in which AI will be deployed, and we will decide how it can best be used.
- The technology is very powerful. Good guys will use it. But so will bad guys.
- The only solution will come not from throwing up our hands and declaring the problem too complex to manage. It will come from a careful, broad and broadly socialised discussion about the kind of society in which we want to live.

A Robust Discussion

A partial listing of some of the milestones and processes undertaken by civil society, international organisations and private companies

One Hundred Year Study on Artificial Intelligence. Founded in 2014, based at Stanford. https://ai100.stanford.edu/about

Research Priorities for Robust and Beneficial Artificial Intelligence. 2015 Open Letter by Future of Life Institute, now signed by over 8000 people. https://futureoflife.org/ai-open-letter/

Global Initiative on Ethics of Autonomous and Intelligent Systems. Launched in 2016 as document on Ethically Aligned Design, leading to 11 standards projects. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_about_us.pdf

Partnership on AI. An initiative of AI researchers from major companies and NGOs founded in late 2016. https://www.partnershiponai.org/about/

UN Group of Governmental Experts on Lethal and Autonomous Weapons Systems. Established in 2016. https://www.unog.ch/80256EE600585943/(httpPages)/7C335E71DFCB29D1C1258243003E8724

Continued on the next page

[33] The idea of proactive safety guidelines that would teach robots to maximize human agency and wellbeing is explored by Christoph Salge and Daniel Polani, "Empowerment as Replacement for the Three Laws of Robotics," *Frontiers in Robotics and AI* 4 (2017): 25. The idea is that AI may think around itself negative constraints, but maximisation of positive values may be more robust to error or circumvention.

[34] For a good overview of the economic impact of cybercrime, see James Lewis, *Economic Impact of Cybercrime — No Slowing Down* (Santa Clara: McAfee, 2018).

[35] The comparison is made explicitly by Igor Linkov, Benjamin D. Trump, Kelsey Poinsatte-Jones and Marie-Valentine Florin, "Governance Strategies for a Sustainable Digital World," *Sustainability* 10, no. 2 (2018): 440.

[36] This argumentation is taken from Paul Hofheinz, *The Ethics of Artificial Intelligence: How AI Can End Discrimination and Make the World a Smarter, Better Place* (Brussels: Lisbon Council, 2018).

Different catchphrases are being used to describe what a broad approach to responsible AI might look like, including "human-centred AI," "AI for good" or "AI for humanity."[37] The previous pages have outlined the various risks AI poses to human agency. But ongoing discussions give cause for confidence — the adoption of AI will force us to more closely evaluate questions of bias, ethics, human rights, safety and security. We may very well end up in a better place than we started.

Five horizontal areas of work can help move us in that direction:

1. Above all, **collaborate**. NGOs, large companies and researchers have all played significant roles in developing this discussion over the last decade. The open nature of this dialogue needs to continue, and government decision making needs to be informed by these broader discussions. Such collaboration is also important to avoiding a race to the bottom in safety, security and ethics.[38]

   All participants in this discussion should aim to maintain the current transparency and openness of AI research and develop a global community of practice with a strong professional ethos. This also means avoiding AI nationalism,[39] keeping research, standards and regulations open, international and globally compatible. In particular, we should be at pains to pull researchers, policymakers and companies from outside the narrow circle of wealthy economies into the discussion. China is particularly important here. Every kind of link, whether intergovernmental, academic or commercial, is essential to developing our future ecosystem of products and services.

   The best knowledge about how AI works comes from using it yourself. If only for this reason, government needs to embrace the use of AI tools in its internal organisation.[40]

2. **Don't throttle development.** For all the risks and concerns outlined in this policy brief, calls to outright ban certain types of research or products (e.g. autonomous weapons) are unlikely to work. Even worse, they will allow those without scruples to plough forward. Every one of the technologies AI enables — autonomous weapons included — brings significant benefits. Rather, the world's most advanced economies should pursue an enthusiastic embrace of AI while addressing specific risks.

   In spite of these measures, and with the best of intentions, accidents will happen. We will need to learn from them and adapt quickly, but we should use the attention on individual safety and security incidents today to build support for a broader agenda.

Continued from previous page

Statement on artificial intelligence by the European Group on Ethics in Science and New Technologies. March 2018. https://ec.europa.eu/jrc/communities/community/humaint/useful-link/statement-artificial-intelligence-european-group-ethics-science-and

Toronto Declaration on Protecting the rights to equality and non-discrimination in machine learning systems, prepared by Amnesty International and Access Now. Adopted May 16, 2018. https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/

Google AI Principles. June 7, 2018. https://www.blog.google/technology/ai/ai-principles/

EU High-level expert group on Artificial Intelligence, producing recommendations by late 2018. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Council of Europe Committee of experts on Human Rights Dimensions of automated data processing and different forms of artificial intelligence. Ministerial recommendations and study expected end 2019. https://www.coe.int/en/web/freedom-expression/msi-aut

[37] The first of these is a research direction, including as an eponymous research initiative at Stanford University (https://hai.stanford.edu), the second of these a United Nations Platform (https://ai4good.org), the third is the title of the French national AI strategy by Cédric Villani, "AI for Humanity: French Strategy for Artificial Intelligence" (Paris: 29 March 2018).

[38] The European Parliament's research report on this matter calls for a global charter for AI. See Peter Bentley, "The Three Laws of Artificial Intelligence: Dispelling Common Myths," in *Should We Fear Artificial Intelligence* (Brussels: European Parliament Research Service, 2018).

[39] Ian Hogarth, "AI Nationalism," Blog Entry, 13 June 2018. https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism.

[40] This argument is emphasized by the 100 year study on AI, "Artificial Intelligence and Life in 2030," *One Hundred Year Study on AI: Report of the 2030 Study Panel* (Stanford: Stanford University, 2016).

3. **Promote awareness and transparency.** The limitations of AI and the security risks opened up through greater automation are real. Awareness of AI risks should also feed into a greater focus on designing more secure systems. Consumers, manufacturers, service providers and policymakers all have a role to play in building demand for more security.

   A number of specific steps can help:

   - Develop new standards and codes of conduct. A wide range of processes are currently underway in this direction, including by individual companies, technical standards bodies and international organisations.
   - Employ the full toolbox for managing AI fallibility, including availability of data, explainability, understandability and testing.
   - Promote labels, certifications and other information mechanisms that help users make a choice for well-designed AI systems.
   - Undertake thorough testing of AI and only employ mission-critical AI applications in the environments they have been tested in.
   - Strategise: Companies and governments need to include sections on safety, security and ethics in their AI strategies, and security, risk management and sustainability strategies should look closely at AI.
   - Observatories formed to monitor AI developments, whether independently or attached to international organisations, can help promote transparency and closer cross-disciplinary cooperation.

   Researchers, policymakers, companies and journalists all have a duty to refrain from AI-related sensationalism and aim to demystify this technology.[41]

4. **Move from awareness to responsibility.** As our concrete norms around liability and responsibility evolve, we need to ensure that ultimate responsibility for AI does not dissipate into vague notions of distributed agency. Machines still don’t make decisions, even if their algorithms do.

   A wide variety of mechanisms can be used to create responsibility, ranging from public and reputational pressures to new regulations and updates to liability rules. The important goal to bear in mind is that it should be relatively clear how responsibility is ultimately attributed to the various human actors involved — both to help individual users protect themselves and aid the development of a commercial ecosystem of insurance, liability and court practice around the use of autonomous agents.

   Regulatory adjustments are likely to proceed through a combination of horizontal and sector specific regulation going into granular detail.

[41] This argument is made by Margaret Margaret, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman et al in “Principles of Robotics: Regulating Robots in the Real World,” *Connection Science* 29, No. 2 (2017): 124-129.

5. **Invest in safe and robust AI.** Support the development of techniques and practices that lead to more secure and robust AI. While no technological solution alone will address the concerns raised in this paper, new technology can tip the balance in favour of control. This is an area for public funding — safe, secure and ethical AI is an underlying public interest and competitive pressures may fail to produce sufficient investment.

   Some focal areas for further research can be noted here:

   - Means for robust testing of AI systems in realistic environments.
   - Tools to provide transparency and explainability to users. Even *ex post* explainability is a must, allowing investigators to piece together "what happened and why" after a decision or an incident.
   - Build data integrity into systems. Robotic systems interacting with the physical world need to be sure that their sensor input and computing have not been compromised. Similarly, data integrity is key to ensuring the evidentiary value of any data, including of logs in case of incidents.
   - Develop means for humans to take over responsibility from automated systems, such as kill switches, as well as coding AI to recognise situations where it should cede control to humans.

10. Epilogue

"The Sorcerer's Apprentice" (a poem of Johann Wolfgang von Goethe's and the eponymous scene in Disney's "Fantasia"), tells the story of a sorcerer who leaves his apprentice (Mickey Mouse in the film) with chores to perform. Tired of fetching water by pail, the apprentice enchants a broom to do the work for him — using magic in which he is not yet fully trained. The floor is soon awash with water, and the apprentice realises that he cannot stop the broom because he does not know how.[42]

The apprentice splits the broom in two with an axe — but each of the pieces becomes a whole new broom that takes up a pail and continues fetching water, now at twice the speed. When all seems lost, the old sorcerer returns and quickly breaks the spell. The poem finishes with the old sorcerer's statement that powerful spirits should only be called by the master himself.

1882 Illustration by Ferdinand Barth

What for centuries seemed like magic is now rapidly entering the realm of the very real and possible. We now face the challenge of making ourselves the master of this magic and avoiding the fate of the apprentice.

[42] Inspiration for this reference comes from Paul Scharre's excellent book, which provides perhaps the best lay-accessible book length exposition on the topics touched upon here, albeit with more of a focus on autonomous weapons and warfare. Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: WW Norton & Company, 2018).

Author

Luukas Ilves is deputy director and senior fellow of the Lisbon Council, a Brussels-based think tank. He has previously served in the Estonian government and European Commission, where he has held various positions responsible for digital policy.

A special thanks to Corinne Cath-Speth, Charina Chou, Luciano Floridi, Ian Goodfellow, Paul Hofheinz, Mikko Hypponen, Toomas Hendrik Ilves, Klen Jäärats, Marten Kaevats, Stéphanie Lepczynski, Mari-Liis Lind, Maarja Naagel, Jan Neutze, Riina Raudne, Kriss-Elin Rokk, Martin Ruubel, Paul Scharre, Siim Sikkut, Karmen Turk and Fernanda Viégas.

Bibliography

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety," *Google Research*, 2016

Asimov, Isaac. *I, Robot* (New York: Gnome, 1950)

Baum, Seth. "On the Promotion of Safe and Socially Beneficial Artificial Intelligence," *AI & SOCIETY* 32, No. 4 (2017): 543-551

Baum, Seth. "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy," *Global Catastrophic Risk Institute Working Paper 17-1*, 2017

Bentley, Peter. "The Three Laws of Artificial Intelligence: Dispelling Common Myths," in *Should We Fear Artificial Intelligence* (Brussels: European Parliament Research Service, 2018)

Boddington, Paula. *Towards a Code of Ethics for Artificial Intelligence* (Springer, 2017)

Boden, Margaret, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman et al. "Principles of Robotics: Regulating Robots in the Real World," *Connection Science* 29, No. 2 (2017): 124-129

Bostrom, Nick. *Superintelligence* (Paris: Dunod 2017)

Buczak, Anna L., and Erhan Guven. "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys and Tutorials* 18, No. 2 (2016): 1153-1176

Carlini, Nicholas, and David Wagner. "Adversarial Examples are Not Easily Detected: Bypassing Ten Detection Methods," *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017

-------. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," *arXiv preprint arXiv:1801.01944*, 2018

Cerf, Vint, Patrick S. Ryan, Max Senger and Richard S. Whitt. "IoT Safety and Security as Shared Responsibility," *Business Informatics* No. 1, (2016): 35

Datta, Amit, Anupam Datta, Ariel D. Procaccia, and Yair Zick. "Influence in Classification via Cooperative Game Theory," *International Joint Conferences on Artificial Intelligence*, pp. 511-517, 2015

Elish, Madeleine Clare. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," *We Robot 2016* (2016)

European Court of Human Rights (Grand Chamber). *Delfi v Estonia: Judgment (Application no. 64569/09),* (Strasbourg: European Court of Human Rights, 2015)

European Commission. *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe* (Brussels: European Commission, 2018)

Future of Life Institute. *Asilomar AI Principles* (Asilomar: Future of Life Institute, 2017). Available at https://futureoflife.org/ai-principles/

Hofheinz, Paul. *The Ethics of Artificial Intelligence: How AI Can End Discrimination and Make the World a Smarter, Better Place* (Brussels: Lisbon Council, 2018)

Hogarth, Ian. "AI Nationalism," *Blog Entry*, 13 June 2018

Huang, Sandy, Nicolas Papernot, Ian Goodfellow, Yan Duan and Pieter Abbeel. "Adversarial Attacks on Neural Network Policies." *arXiv preprint arXiv:1702.02284*, 2017

Huang, Xiaowei, Marta Kwiatkowska, Sen Wang and Min Wu. "Safety Verification of Deep Neural Networks," *International Conference on Computer Aided Verification*, pp. 3-29, 2017

KPMG International. *Autonomous Vehicles Readiness Index* (Geneva: KPMG International, 2018).

Knight, Will. "Google's Self-Training AI Turns Coders into Machine-Learning Masters," *Technology Review*, 17.01.2018.

Lewis, James. *Economic Impact of Cybercrime — No Slowing Down* (Santa Clara: McAfee, 2018)

Linkov, Igor, Benjamin D. Trump, Kelsey Poinsatte-Jones,and Marie-Valentine Florin. "Governance Strategies for a Sustainable Digital World," *Sustainability* 10, No. 2 (2018): 440

Liu, Qiang, Pan Li, Wentao Zhao, Wei Cai, Shui Yu and Victor CM Leung. "A Survey on Security Threats and Defensive Techniques of Machine Learning: a Data Driven View" *IEEE access* 6 (2018): 12103-12117

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv preprint arXiv:1706.06083*, 2017

Mahoney, Donna. "Flying on Autopilot Improves Airlines Safety but Can Lead to Errors," *Business Insurance*, 28 February 2016

Malliaraki, Eirini. "Toward Ethical Transparent and Fair AI/ML: A Critical Reading List," *Blog*, 2018.

Miller, Simon, Christian Wagner, Uwe Aickelin, and Jonathan M. Garibaldi. "Modelling Cyber-Security Experts' Decision Making Processes Using Aggregation Operators," *Computers and Security* 62 (2016): 229-245

Milli, Smitha, Dylan Hadfield-Menell, Anca Dragan and Stuart Russell. "Should Robots Be Obedient?" *arXiv preprint arXiv:1705.09990*, 2017

Müller, Vincent C. "Autonomous Killer Robots are Probably Good News," *Drones and Responsibility*, pp. 77-91, 2016

Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion" in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Berlin: Springer, 2016)

Nordrum, Amy. "Popular Internet of Things Forecast of 50 Billion Devices by 2020 is Outdated," *IEEE spectrum* 18, 2016

Nunes, Eric, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart and Paulo Shakarian. "Darknet and Deepnet Mining for Proactive Cybersecurity Threat Intelligence," *arXiv preprint arXiv:1607.08583*, 2016

Nunes, Eric, Paulo Shakarian, Gerardo I. Simari and Andrew Ruef. "Argumentation Models for Cyber Attribution," *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 837-844, 2016

Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha and Michael Wellman. "Towards the Science of Security and Privacy in Machine Learning," *arXiv preprint arXiv:1611.03814*, 2016

Philbeck, Thomas, Nicholas Davis and Anne Marie Engtoft Larsen. *Value, Ethics and Innovation: Rethinking Technological Development in the Fourth Industrial Revolution* (Geneva: World Economic Forum, 2018)

Reese, Hope. "Why Microsoft's 'Tay' AI bot went wrong." *Tech Republic*, 24 March 2016

Russell, Stuart, Daniel Dewey and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine* 36, No. 4 (2015): 105-114

Salge, Christoph, and Daniel Polani. "Empowerment as Replacement for the Three Laws of Robotics," *Frontiers in Robotics and AI* 4 (2017): 25

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War* (New York: WW Norton & Company, 2018)

Scherer, Matthew U. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies," *Harvard Journal of Law and Technology* 29 (2015): 353

Schwab, Klaus. *The Fourth Industrial Revolution* (New York: Crown Business, 2017)

Schwab, Klaus, and Nicholas Davis. *Shaping the Fourth Industrial Revolution* (World Economic Forum, 2018)

Smith, Brad. "Facial Recognition Technology: The Need for Public Regulation and Corporate Responsibility," *Microsoft Blog*, 3 July 2018

Tian, Yuchi, Kexin Pei, Suman Jana,and Baishakhi Ray. "Deeptest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars," *Proceedings of the 40th International Conference on Software Engineering*, pp. 303-314, 2018

Townsend, Kevin. "IBM Describes AI-powered Malware That Can Hide Inside Benign Applications," *Security Week*, 13 August 2018

Turk, Karmen, and Maarja Pild. *Analüüs SAE Taseme 4 ja 5 Sõidukite Kasutusele Võtmiseks: Kitsas ja Lai Vaade* (Tallinn: Triniti, 2017)

One Hundred Year Study on AI. "Artificial Intelligence and Life in 2030," *One Hundred Year Study on AI: Report of the 2030 Study Panel* (Stanford: Stanford University, 2016)

Price, W Nicholson II. "Regulating Black-Box Medicine," *Michigan Law Review* 116 (2017): 421

Vattapparamban, Edwin, İsmail Güvenç, Ali İ. Yurekli, Kemal Akkaya and Selçuk Uluağaç. "Drones for Smart Cities: Issues in Cybersecurity, Privacy and Public Safety," *Wireless Communications and Mobile computing Conference*, 2016

Villani, Cédric. *AI for Humanity: French Strategy for Artificial Intelligence* (Paris: 29 March 2018)

Wagstaff, Kiri L. "Machine Learning That Matters," *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1851-1856, 2012

Wong, Joon Ian, and Nikhil Sonnad. "Google's AI Won the Game Go by Defying Millennia of Basic Human Instinct," *QZ.com*, 25 March 2016

Wooldridge, Michael, and Nicholas R. Jennings. "Intelligent Agents: Theory and Practice," *The Knowledge Engineering Review* 10, No. 2 (1995): 115-152

Wu, Desheng, David L. Olson and Alexandre Dolgui. "Artificial Intelligence in Engineering Risk Analytics," *Engineering Applications of Artificial Intelligence* 65 (2017): 433-435

Yampolskiy, Roman V. "Taxonomy of Pathways to Dangerous Artificial Intelligence." *AAAI Workshop: AI, Ethics, and Society*, 2016

Yampolskiy, Roman V., and M. S. Spellchecker. "Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures," *arXiv preprint arXiv:1610.07997*, 2016

Yampolskiy, Roman, and Joshua Fox. "Safety Engineering for Artificial General Intelligence," *Topoi* 32.2 (2013): 217-226

Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk," *Global Catastrophic Risks* 1, No. 303 (2008): 184