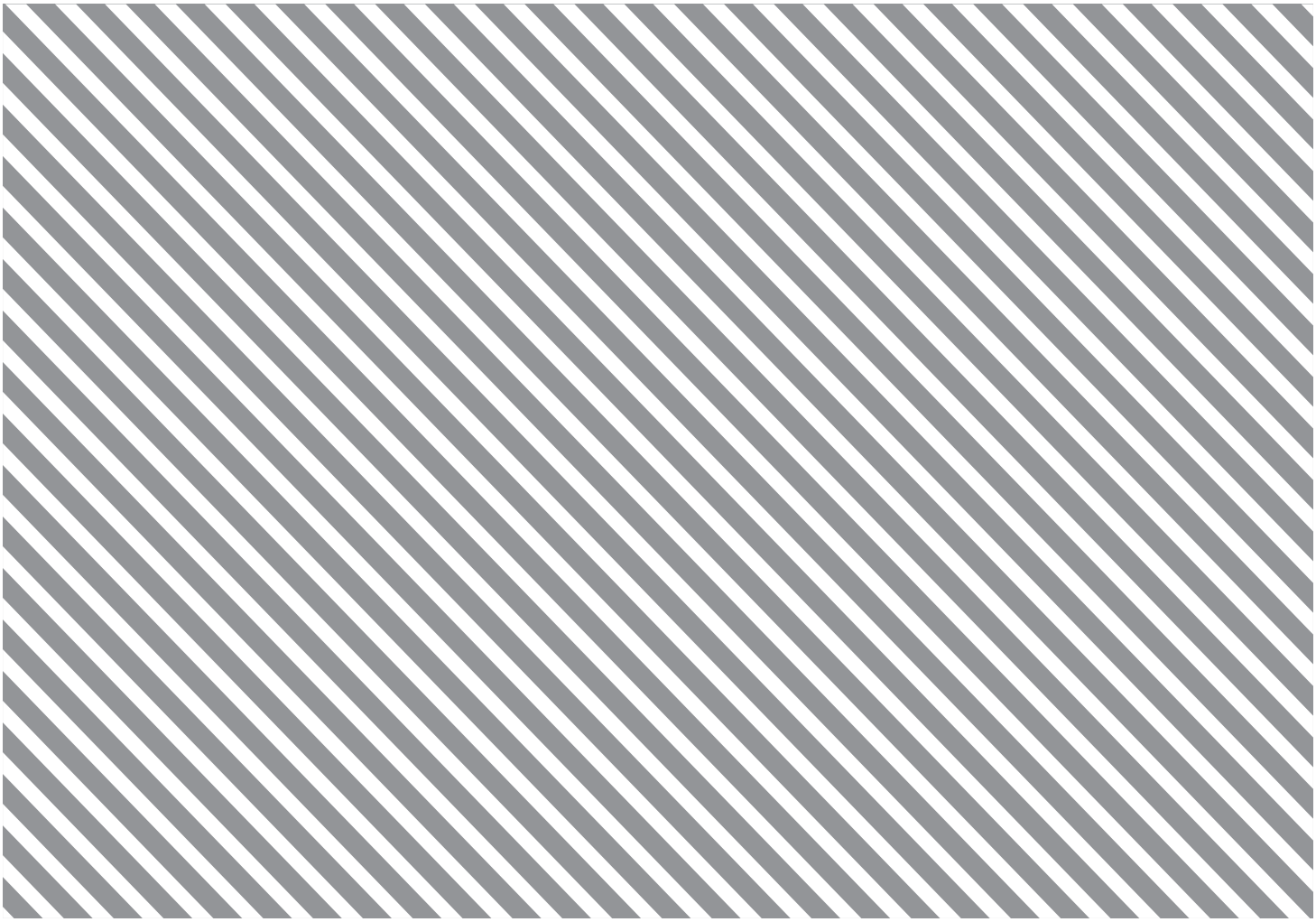
WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

White Paper

AI Governance
A Holistic Approach to Implement Ethics into AI

January 2019



World Economic Forum
91-93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland
Tel.: +41 (0)22 869 1212
Fax: +41 (0)22 786 2744
Email: contact@weforum.org
www.weforum.org

A longer version of this paper will appear in the *Duke Law & Technology Review*

Contents

Introduction

This White Paper aims to enrich the ongoing debate about implementing ethical considerations into artificial intelligence (AI) by looking at possible means and mechanisms to apply ethical values and principles in AI-driven technology and machines in order to contribute to building a human-centric AI-society.[1] The goal is to outline approaches to determine an AI governance regime that fosters the benefits of AI while considering the relevant risks that arise from the use of AI and autonomous systems. To this end, various concepts that could be applied to ensure that the use of AI does not conflict with ethical values are posited. The first section of this paper reviews certain ethical concerns encountered with the use of AI. The second section outlines and discusses the advantages and disadvantages of different governance instruments that could be evoked to implement ethics in AI applications. The third section presents various practical approaches for the governance of AI applications. Based on these insights, the fourth section concludes with recommendations to develop a holistic AI governance regime.

I. Ethical concerns in AI applications

1. Definition of basic terms

a. AI

Despite intense ongoing discussion on the possible regulation of AI, no unanimous agreement on the definition of AI exists.[2] AI as a term was first coined by John McCarthy for the Dartmouth Summer Research Project of 1956.[3] McCarthy defined AI as a machine that behaves "in ways that would be called intelligent if a human were so behaving."[4] This definition, however, does not mention the technical functionality of AI. Focusing more on a technology's ability to adapt to changing circumstances, a further definition of AI refers to "technology (software, algorithm, a set of processes, a robot, etc.) that is able to function appropriately with foresight of its environment".[5] The UK Government Office for Science defines AI as "the analysis of data to model some aspect of the world. Inferences from these models are then used to predict and anticipate possible future events".[6] This involves the creation of statistical models that use series of algorithms, or step-by-step instructions that computers can follow to perform a particular task.[7]

Technically, AI is, in the main, empowered by machine learning algorithms, i.e. algorithms that change in response to their own received inputs and consequently improve with experience.[8] Machine learning has to be distinguished from deep learning. Deep learning algorithms consist of several non-linearly connected layers (so-called neural networks) where each unit in the bottom layer takes in external data, such as pixels of images for the purpose of face recognition systems, then distributes that information up to some or all of the units in the next layer. Each unit in that second layer then integrates its inputs from the first layer, using a simple mathematical rule, and passes the result further up to the units of the next layer.[9] The input data accordingly passes through numerous layers of statistical data operations to produce the requested output data. Based on statistical techniques, such output is – as is the case for all AI-generated output – probabilistic in nature.[10] In view of the different layers being non-linearly connected with each other in the form of neural networks, corresponding deep learning based processes become so complex that their decision-making processes become entirely opaque and, therefore, decisions ultimately taken by such systems cannot be understood by humans (the so-called black box effect).[11] The multilayered approach allows corresponding machines to not only follow pre-programmed decisions but to respond to changes within their environment. A further example in addition to face recognition systems referred to above are autonomous cars that can make real-time decisions about speed and direction by administering sensor-based data without input from a human user.[12] In a summary, therefore, AI can be described as a technology that is able to adapt itself to changing circumstances on the basis of a certain self-learning ability and produce specific output independent of human control.

b. Ethics

Ethics is commonly referred to as the study of morality.[13] "Morality" for the purpose of this paper is understood as a system of rules and values for guiding human conduct and principles for evaluating those rules.[14] Consequently, ethical behaviour does not necessarily mean "good" behaviour. Instead, it indicates compliance with specific values.[15] Such values are commonly accepted as being part of human nature (e.g. the protection of human life, freedom and human dignity) or as a moral expectation characterizing beliefs and convictions of specific groups of people (e.g. religious rules). Moral expectations may also be of individual nature (e.g. an entrepreneur's expectation that employees accept a company specific code of conduct). This broad definition is used here as the intention of this article is not to approach AI from a specific normative perspective or to analyse AI in a moral sense but to contribute to the discussion on the determination of appropriate regulatory means to implement ethics into AI. In addition, the benefit of this neutral definition of ethics is that it addresses the issue of ethical diversity from a regulatory and policy-making perspective.

2. Potential benefits of AI applications

A recent study, conducted on behalf of the European Parliament, concludes that AI applications will likely be used in almost all aspects of our daily lives.[16] AI's benefits include the reduction of economic inefficiencies and labour costs, as well as an increase in high-skilled jobs. Moreover, AI can help companies to understand their individual customers better and thus develop more customized products tailored to their specific needs. The increasing flexibility of smart factories is likely to play a decisive role in this regard.[17] Knowing the customer better also results in more individualized and, as a consequence, economically efficient sales and marketing strategies.[18] While these benefits appear to favour companies in modern economic systems, AI applications can bring specific benefits to consumers. These, however, predominantly depend on where and how AI is to be applied. For example, in the individualization of the manufacturing process, one benefit to consumers is that the variety of products offered to them increases. The increasing flexibility of smart factories also multiplies competition between companies that might not have been considered as competitors.[19] This increased competition can force companies to pass on the reduced AI-driven production costs to their customers so they benefit from cheaper prices.

3. Potential risks of AI applications

a. Loss of jobs

Technological change has traditionally been accompanied by fundamental societal changes, often including massive job losses.[20] Historically, for instance, with the completion of the first US transcontinental telegraph line in 1861, the services of Pony Express riders became obsolete.[21] Telegraph lines, however, soon became the basic fundament for the emergence of the new telecommunication industry, creating myriad new jobs over time. The increasing use of AI poses the question of whether it can be seen as the new telegraph line, forming a breeding ground for a new job-intensive AI-industry, or whether the delegation of more tasks to AI systems may lead to a significant number of job losses, even in the long term.[22] This raises uncertainties about whether a more automated, digital society and economy will provide sufficient opportunities for people to earn a living.[23] While precise calculations are still lacking, studies conducted so far estimate that 49% of activities performed in jobs,[24] or between 21% and 38%[25] of jobs in the developed world, could be lost as a result of an increasingly digitalized and automated economy. A recent study conducted in the United Kingdom estimates that countervailing displacement and income effects are likely to broadly balance each other out over the next 20 years.[26]

b. Liability for damages caused by AI systems

AI systems are increasingly being used in close proximity to humans, which raises the question of who should be held liable for potential damages caused by their operation.[27] This is even more relevant as a malfunction in automated systems could have multiplying effects.

The critical ethical issue in this respect is whether a human being is responsible for damages caused by an AI-driven or otherwise automated machine which, after consideration of certain data, has taken an autonomous decision and caused harm to a human's life, health or property. While one could argue that the person having implemented or used the AI system in fulfilment of an owner obligation is responsible, this question will become more critical as more and more autonomous decisions are made by AI systems. Legal accountability is generally not found if independent events or decisions cause specific damage, unless the law provides for strict liability regimes, as is the case in Europe in product liability law.[28] Fault-based liability regimes might therefore expose victims of AI-caused damages to significant protection gaps.

Whether the existing mixture of fault-based damages compensation regimes and strict liability rules on product liability are appropriate for potential harm caused by AI and autonomous systems is subject to debate.[29] Responsibility, accountability and liability are some of the fundamental ethical concerns that must be discussed in depth in relation to new AI applications.[30]

c. Lack of transparency of AI

AI systems are often criticized for their lack of transparency.[31] The UK Information Commissioner's Office expressly states: "The complexity of the processing of data through such massive networks creates a 'black box' effect. This causes an inevitable opacity that makes it very difficult to understand the reasons for decisions made as a result of deep learning."[32] Transparency is required for various reasons:[33] from a user perspective, transparency is important to build trust in the use of an AI system. Users need to understand what an AI system will do in different circumstances. AI systems should therefore not behave in an unexpected manner.[34]

d. Loss of humanity in social relationships and lack of protection of human life and human dignity

Ever more critical, AI has the potential to cause fundamental changes to humanity: "What is changing in our young, fast-growing digital civilisation is that we can delegate decisions in our individual, family or social lives to technology. Human existence can be subcontracted to software. (…) We've already started putting aside our feelings, intuitions and dreams in favor of more reasonable choices, calculated by an algorithm and powered by objective data."[35] In addition, more automation and reliance on AI to make decisions in our daily lives may lead to a decrease in social contacts. Indeed, AI applications, such as healthcare robots in hospitals, service robots for elderly people or service robots used in the field of tourism and, last but not least, AI-enabled toys, may result in increasing man-to-machine interaction. It is unclear how this development – more human–machine interaction, on the one hand, and fewer social contacts, on the other – may affect our emotional life and ways of thinking.[36] Even typical human strengths, such as emotions and intuition, could be affected significantly by the increasing reliance on AI for decision-making purposes.[37] The new technological developments pertaining to the implementation and use of AI will consequently give rise to fundamental questions about what human life is, what humanity is, what human life and dignity mean and what the relationship to AI systems are when it comes to social interaction with corresponding machines. Further issues regarding AI systems that are used for social interaction are how such systems should behave from an ethical and moral point of view and to what extent self-learning mechanisms and autonomous behaviour should be allowed.[38] A world in which computers can "fake" human emotion and AI can be used to produce fake news and information is difficult for humans to navigate. Simply put, humans are not equipped to live in a world where they are asked to constantly judge what is false and what is real.

e. Loss of privacy

To make intelligent decisions, AI systems need to collect and process data. Thus, the access to data is fundamentally important to the development of digital technologies in general, and AI in particular.[39] With regard to personal data, however, a major concern in certain societies is to ensure that the privacy of such data is protected and maintained.[40] In some societies, it is considered crucial to make sure that, while accessibility of non-personal data is improved, sufficient data protection standards are implemented in relation to personal data.[41] From a European perspective, the General Data Protection Regulation, a new and stricter regulatory framework, came into force on 25 May 2018.[42] At the same time, appropriate means and mechanisms must be implemented to protect AI systems against abuse. That need is underlined by the ongoing discussion about whether an automobile infotainment system susceptible to being hacked can lead to the liability of the car manufacturer.[43]

f. Loss of personal autonomy

While the development of intelligent assistants may be convenient and help to manage administrative and other daily tasks, in certain respects the rise of intelligence and autonomy in machines and software tools may also decrease the intelligence and autonomy of the human user. Digital dementia is a phenomenon described by psychologists as a possible consequence of overusing digital technology, which could result in the deterioration or breakdown of cognitive abilities.[44] Overuse of digital technology may also have an impact on personal autonomy, depending on the degree of digital assistance that is increasingly relied upon to complete even trivial tasks, such as watering indoor plants.[45] As a consequence of the growing reliance on digital assistance, basic human capabilities could be lost.[46] Indeed the self-determination theory suggests that humans need autonomy to function properly in life.

g. Restriction in the plurality of opinions and competition – the information bias of AI and autonomous systems

A further critical issue is that AI applications reflect the background and bias of the source that programmed them.[47] In view of the rapid development of digital products and markets, such bias multiplies quickly and consequently may have a widespread impact.[48] The increasing use of algorithms can further reduce the plurality of views expressed in public discussions. The use of chat bots is an example in this regard. Chat bots pick up certain views and facts and share them with as many readers as possible. This automated mass distribution may cause critical information bias and distort predominant public opinion. This is of particular concern to society if incorrect or biased facts (often referred to as fake news) are intentionally spread by chat bots to influence certain decision-making processes.[49] Consequently, corresponding new communication strategies may have tortious interference on elections and other democratic decision-making procedures, thereby causing significant concerns.[50] In addition to a possible reduction in the plurality of views and opinions, algorithms may also reduce competition, impacting negatively on innovation.[51]

h. Error proneness and susceptibility to manipulation of AI

Using and implementing AI from a technical perspective means using and implementing software and computer systems. It is acknowledged, however, that both software and hardware do not function correctly all the time; rather, different kinds of errors may occur.[52] Also to be kept in mind is that AI-generated decisions and results are based on algorithms that use statistical models analysing certain amounts of data.[53] But the use of statistical models may generate faulty decisions and results, be it because the data analysed for a specific case does not reflect the individual circumstances of the respective scenario, because the data analysed are biased or incorrect, or because the statistical model is incomplete or incorrect.[54] From a legal perspective, decision-making processes relying on statistical models involve automatic discrimination with regard to cases that differ from the statistical role model.[55]

Further, computer and software technology is susceptible to errors and manipulation.[56] Even computer and software systems believed to be secure, including the network of the Government of the Federal Republic of Germany, have been hacked successfully.[57] The German Federal Office for Information Security (*Bundesamt für Sicherheit in der Informationstechnik*, BSI) in general concluded in its report on the state of IT security in Germany in 2017 that "the risk situation is continuously tense and at a high level". According to the BSI, "vulnerabilities exist in software, and in some cases even hardware products, which are used most often. These vulnerabilities enable attackers to recover information or gain control over systems".[58] This indicates that software and hardware systems, which are at the root of AI, are highly error-prone and susceptible to manipulation. Flash crashes have already provided insights into what might come to pass.

i. Manipulation, surveillance and illegal behaviour

Finally, the risk that AI can be abused for manipulation, surveillance or other quasi-legal purposes is high. For instance, democratic elections could be manipulated,[59] facial recognition systems could be abused to control citizens[60] and companies could use price determination algorithms to set sales prices above the market level, thereby harming consumers.[61]

II. Means to implement ethics in AI applications

The potential benefits and the variety of concerns involved in the use of AI demonstrate the need for a variety of ideas on how to mitigate or even eliminate the risks, so the technology can best be used. Technical solutions and traditional regulatory approaches are considered below, including binding and non-binding measures of self-regulation:

1. Technical means and mechanisms – Ethics compliance by design

a. Bottom-up versus top-down approaches – The Tay example

To implement ethical decision-making criteria technically, both a bottom-up approach and a top-down approach are possible.[62] In a bottom-up approach, machines would be expected to observe human behaviour in specific situations and learn how to make ethical decisions on that basis. However, by observing people, the machines would not adopt what is ethical but what is common.[63] In 2016, shortly after its launch, Microsoft's chat bot Tay started making racist, inflammatory and political statements that it had been taught by users determined to undermine it.[64] Therefore, from a technical perspective, it appears that a top-down approach is better suited to implement ethics into AI. In such an approach, ethical principles would be programmed directly into an AI system.[65]

b. Casuistic approach

Ethical principles could be implemented in AI systems on a casuistic basis. Machines would be programmed to react specifically in each situation where they may have to make an ethical decision. A healthcare robot, for instance, could be programmed to always consider the will of its user, i.e. the patient, before taking specific action. If the user did not express a clear will in relation to a specific situation, the robot would need to ask for the user's confirmation before taking action. In emergency situations, a healthcare robot could be programmed to first check its user's advance directive before initiating first aid measures. The robot could even be programmed to take different decisions depending on the type of emergency and the user's state of health. Difficulties, however, would arise when no advance directive is available and the user is not in a position to express their will any longer. Probably, to protect the absolute fundamental right to human life,[66] in this scenario an AI system default setting should decide on the action that has the highest probability of saving the user's life. But there are fears that a system that teaches itself might decide that it knows better what is needed than its original programmers.

c. Dogmatic approach

Rather than anticipating all possible scenarios in which an AI system would need to take an ethical decision and then programming the AI system, the systems could be programmed in line with a specific ethical school of thought. Examples include utilitarianism, Kantian ethic,[67] Asimov's Three Laws of Robots[68] or the Golden Rule,[69] international philosophies and different religions that propose that one should not treat others in a way that they would not like to be treated.[70] A major issue is that an AI system programmed in line with a certain ethical school of thought would make decisions slavishly on the basis of that specific school. It is, however, not yet clear whether AI systems could be so programmed. But by doing so, a decision could in a specific scenario end up being unethical. Most ethicists apply the rules of various schools of thought to resolve a specific ethical issue in order to make well-balanced decisions.[71] Therefore, it appears – at least for the time being – that the technical approach is preferable to program ethical principles into AI systems on a more casuistic basis relying on specifically programmed decision-making structures. Still, it remains a challenge for AI system designers to generally deal with this question and decide on a design philosophy for algorithmic decision-making frameworks. As a potential approach to resolve the issue of situation-specific ethics applications, the recommendation is for ethical requirements for computational systems to be developed collaboratively and reviewed to achieve consistency in the decision-making process.[72] Close cooperation between researchers, developers and policy-makers is necessary to develop a common understanding of the relevant ethical principles that are culturally relevant on the basis of which the "good AI society" can be developed.[73] The World Economic Forum is working with partners on the concept of an AI "ethics switch", which would allow the AI to change ethical protocols in diverse jurisdictions, or switch off in the event of exceeding its mandate.

d. Implementing AI on a technical meta-level

In view of the autonomous nature of decisions made by AI, an AI-driven monitoring system that controls a machine's compliance with a predetermined set of laws and ethical rules on a meta-level ("guardian AI") could be developed. Such guardian AI could technically interfere in the basic AI's system and directly correct unlawful or unethical decisions. Also, a corresponding guardian AI could be programmed to report the basic AI's unlawful or unethical decision to an appropriate enforcement authority or agency.[74] These requirements and benefits can be transformed in concrete technical solutions, when they are available.[75]

e. The insufficiency of technical means and mechanisms

However, while technical means may eventually be able to resolve ethical issues, these approaches are probably insufficient to make sure that AI systems do indeed take ethical considerations into account in their decision-making process. AI systems are constructed by, programmed by and used by humans and companies. Therefore, unless the people and companies responsible for programming and using them are committed to ethical standards for personal reasons, humans and companies will only program and use AI systems in an ethically aligned manner if they are forced to do so by binding legal rules, or if they believe that a corresponding ethically aligned system design is economically or otherwise beneficial to them. To make sure that AI systems behave according to ethical principles, it is therefore necessary to adopt a variety of agile governance mechanisms, including, for example, binding legal requirements or the creation of economic incentives to promote ethically aligned AI system design.

2. Policy-making instruments

Considering the insufficiency of technical means to ensure ethical AI decision-making processes, it is necessary to consider possible policy-making instruments, such as legislation. While legislation has the advantage of providing binding and enforceable rules that are established and consequently generally accepted on the basis of a democratic process that ensures transparency and the participation of the people and relevant interest groups, laws can often only protect a minimum consensus of ethical rules. Therefore, legislation may not be an appropriate regulatory instrument insofar as the specific ethical interests of selected individuals are concerned. In addition, in view of their territorial limitation, laws only bind people of, and within, respective national states.[76] It is often difficult to respond to technical developments with regulatory mechanisms sufficiently quickly to keep up with the technology. Also, legislation is often perceived as having a negative impact on innovation and may ultimately disadvantage domestic businesses in relation to businesses in less regulated countries. However, legislation can bring about new incentives to innovate as companies compete to adopt compliant technologies and business models. For instance, with regard to data protection, efficient legislation can be considered a competitive advantage and incentivize businesses to develop innovative privacy by design solutions and transfer their registered offices to countries assuring a high level of data protection.[77] This consideration should also be kept in mind in relation to other ethical rules and values. Customers might generally like the fact that businesses are subject to certain strict and binding statutory regulations and accordingly prefer services rendered by those companies that are subject to corresponding strict laws. A balanced governance approach therefore needs to take into account the potential anti- as well as pro-competitive effects of legislative regulation.

Therefore, for governance-making purposes, it is important to be aware of the plurality of policy-making instruments that could be considered, as an alternative or in addition to legislation, for implementing ethical considerations into AI applications. Possible policy-making instruments in addition to legislation are:

- international resolutions and treaties[78]
- bilateral investment treaties (BITs)[79]
- self-regulation and andardization[80]
- certification
- contractual rules
- soft law[81]
- agile governance[82]
- monetary incentives[83]

III. Two practical approaches to implement ethics in AI systems

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative") and the World Economic Forum's project on Artificial Intelligence and Machine Learning are concrete practical approaches to implement ethics into AI and autonomous systems.

1. The IEEE Global Initiative

The IEEE Global Initiative is a programme of the Institute of Electrical and Electronics Engineers ("IEEE") launched in December 2015. A primary goal of the initiative is to ensure that technologists are educated, trained and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems.[84] To this end, the IEEE Global Initiative issued the document, "Ethically Aligned Design – A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems"[85] and, with its standards of the so-called IEEE P7000™ series, presents specific proposals for actual operational standards that can be adopted by AI and autonomous system designers.[86]

The annual "Ethically Aligned Design" reports summarize insights and recommendations of reference to technologists in the related fields of science and technology who are developing and programming AI and autonomous systems. The document, to this end, identifies pertinent issues and "candidate recommendations", which facilitate the emergence of national and global policies that align with these principles.[87] The document, and in particular its recommendations, can be used as a basis for developing operational standards.[88]

2. The World Economic Forum project on Artificial Intelligence and Machine Learning

With its focus on international public-private partnerships, the World Economic Forum provides a neutral and objective platform to help countries and businesses that are struggling with policy implementation and AI governance. The Forum is establishing Centres for the Fourth Industrial Revolution in San Francisco, Tokyo, Beijing and Mumbai. Affiliate Centres are also being planned globally. Governance projects for AI and other technologies will be co-created with governments, businesses, academics and civil society at these Centres. The projects include:

a. Unlocking public-sector AI

Although AI holds the potential to vastly improve government operations, many public institutions are cautious about harnessing it because of concerns over bias, privacy, accountability, transparency and overall complexity. Baseline standards for the public sector's responsible procurement and deployment of AI can help overcome these concerns, opening the door to new ways for governments to better interact with and serve their citizens. Also, as a softer alternative to regulation, governments' significant buying power and public credibility can drive private-sector adoption of these standards.

b. AI Board leadership toolkit

As AI increasingly becomes an imperative for business models across industries, corporate leaders will be required to identify the specific benefits this complex technology can bring to their businesses as well as their concerns about the need to design, develop and deploy it responsibly. A practical set of tools can help Board Members and decision-makers ask the right questions, understand the key trade-offs and meet the needs of diverse stakeholders, as well as consider and optimize certain approaches, such as appointing a Chief Values Officer or creating an Ethics Advisory Board.

c. Generation AI

This project specifically deals with developing standards to protect children. AI is increasingly being imbedded in children's toys, tools and classrooms, creating sophisticated new approaches to education and child development tailored to the specific needs of each user. However, particular precautions must be taken to protect society's most vulnerable members. Actionable guidelines can help address privacy and security concerns arising from data unknowingly collected from children, enable parents to understand the design and values of these algorithms, and prevent biases from AI training data and algorithms undermining educational objectives. Transparency and accountability can build the trust necessary to accelerate the positive social benefits of these technologies for all.

d. Teaching AI ethics

Decisions regarding the responsible design of AI are often made by engineers who receive little training in the complex ethical considerations at play in their designs' various real-world uses. Universities are still struggling to find effective ways to integrate these issues into curricula for technical students. The World Economic Forum Global Future Council on Artificial Intelligence and Robotics is creating a repository of actionable and useful material for faculty who wish to add social inquiry and discourse into their AI coursework.

e. The regulator of the future

Another way of addressing the problem of adequately implementing ethics into AI is to reimagine the regulator as an entity that works with business to certify AI products as well as protect the public.

IV. AI Governance – Determining the appropriate regulation design

Designing an appropriate AI governance regime is a complex challenge, as a careful risk assessment – often referred to as an "impact assessment" – must be conducted. This assessment is particularly complex with regard to the issue of AI governance.

1. The need to conduct a risk assessment with regard to new technologies

New technologies are generally driven by optimistic expectations of the potential benefits that the researchers and developers intend to achieve. Yet new technologies always entail new risks. One illustration is the exploration and development of nuclear power. The optimistic expectation initially was that this new technology would resolve the world's energy supply problem. The consequences that humanity is still facing are the development of nuclear weapons and the as yet unresolved challenge to environmentally and sustainably deal with nuclear waste. So what lesson was learned? Should we not engage in new technologies because of potential abuses and unwanted side effects? More concretely, should the fear of an autonomous combat robot and other potentially uncontrollable AI systems stop us from using AI? From a realistic point of view, this question can only be answered in the negative. At the same time, however, the lessons learned from history teach the need to be cautious and to assess potential risk scenarios carefully before implementing and establishing a potentially risky and uncontrollable new technology.[89] On this basis, abuse and risk prevention means and mechanisms can be employed. A corresponding risk assessment and scientific review involving relevant experts and concerned people may even result in the definition of use cases that show the circumstances in which a certain technology like AI should not be employed at all. For other use cases, specific preconditions, such as the need to pursue marketing authorization procedures or implement specific security technologies, must be considered.

Obviously, this may result in additional regulation and corresponding law enforcement actions. However, this process and the regulation that may ultimately be found to be appropriate as a consequence of such risk assessment should be considered as the *conditio sine qua non* of advancing towards a more digitalized and automated living and working environment while avoiding opening a Pandora's box. In addition, conducting risk–benefit assessments and implementing risk and abuse prevention mechanisms not only protect people and their fundamental rights but further increase the general acceptance of new technologies and thus ultimately result in economic welfare gains.

2. The complexity of AI governance

In relation to AI, designing an appropriate governance system is particularly difficult – first, because of the diverse nature of ethical concerns; second, due to the difficulty of determining the appropriate regulatory instrument; and, third, because of the complex interactions between the relevant technology, the economy and markets, individual humans and society, as well as the environment and, ultimately, politics and regulation.

a. Ethical diversity

While the focus has long been on high-speed Internet access, debate is now addressing the urgent topic of the ethical and societal implications that the digital transformation in general and AI in particular is likely to have. The European Commission's Group on Ethics in Science and New Technologies has presented a comprehensive list of ethical concerns, which are summarized in Figure 1.[90]

Figure 1: The ethical principles of the Group on Ethics in Science and New Technologies

The ethical principles of the European Group on Ethics in Science and New Technology		
Human dignity	Justice, equity and solidarity	Security, safety, bodily and mental integrity
Autonomy	Democracy	Data protection and privacy
Responsibility	Rule of law and accountability	Sustainability

Source: European Group on Ethics in Science and New Technologies, "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems" (2018)

The purpose of this paper is not to discuss the content-related details pertaining to the ethical principles that might be incorporated by AI applications, which requires a broader and separate debate across national, religious and cultural boundaries. What is particularly relevant for the topic dealt with herein is the existing variety and diversity of ethical values, their priorities and relationship between them.

It goes without saying that there are fundamental and universal concerns as defined, for instance, in Art. 2 of the Treaty on European Union: "The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance,

justice, solidarity and equality between women and men prevail." Fundamental human values are further set out in the UN's Universal Declaration of Human Rights and other declarations that expand these rights to specific groups, such as children.[91] In contrast, other ethical concerns reflecting the beliefs of certain individual convictions or communities of values should only be regulated in a manner that reflects the voluntary nature of ethical compliance. This diversity of values needs to be taken into consideration as regards the possible regulation of ethics. That is all the more true in view of the fact that even fundamental needs, for example the protection of human dignity, may be accepted on a global basis but remain controversial when they are brought to life by defining specific requirements and duties to be complied with by concrete AI applications.

An assessment of the ethical implications of AI applications also strongly depends on the relevant cultural and economic framework conditions. This becomes particularly apparent, for instance, in the field of education and is addressed in the work of the World Economic Forum Teaching Ethical AI project. Taking the example of the German Federal Network Agency's decision regarding Cayla,[92] from a US and European perspective that decision will generally be considered ethically justified in view of the need to protect a child's right to privacy. As more of these devices come onto the market, often marketed as educational toys, the questions that arise around the ethics of AI are writ large in this microcosm. Privacy, bias, surveillance, manipulation, democracy, transparency, accountability – all can be challenged by the AI-enabled toy. However, an ethical evaluation may be different from the perspective of developing countries. In many of them, being able to speed up and increase access to education is believed by most economists as the best way to close the gap between the developed and developing world. AI-enabled toys might one day achieve precision education (education using AI that teaches each child individually) and, as our children will be working with autonomous robots, this may be excellent preparation. The difficult question for regulators then is how those potentially good outcomes will be balanced, in particular considering possible additional obligations that may arise for relevant AI companies. For instance, if a regulator should infer that AI-enabled toys may be offered for educational purposes and thus the relevant AI company collects the children's data, even if it is not being stored, should the company red flag children who share suicidal thoughts, other self-harming behaviour or threat scenarios? Ethically, one could argue that technology enables a company to protect a child's life by informing the parents of possible dangerous scenarios. Whether privacy and private autonomy or the protection of a child's health and life is attached greater weight, however, will most likely not be decided unanimously across the globe.

b. Selecting the appropriate regulatory instrument

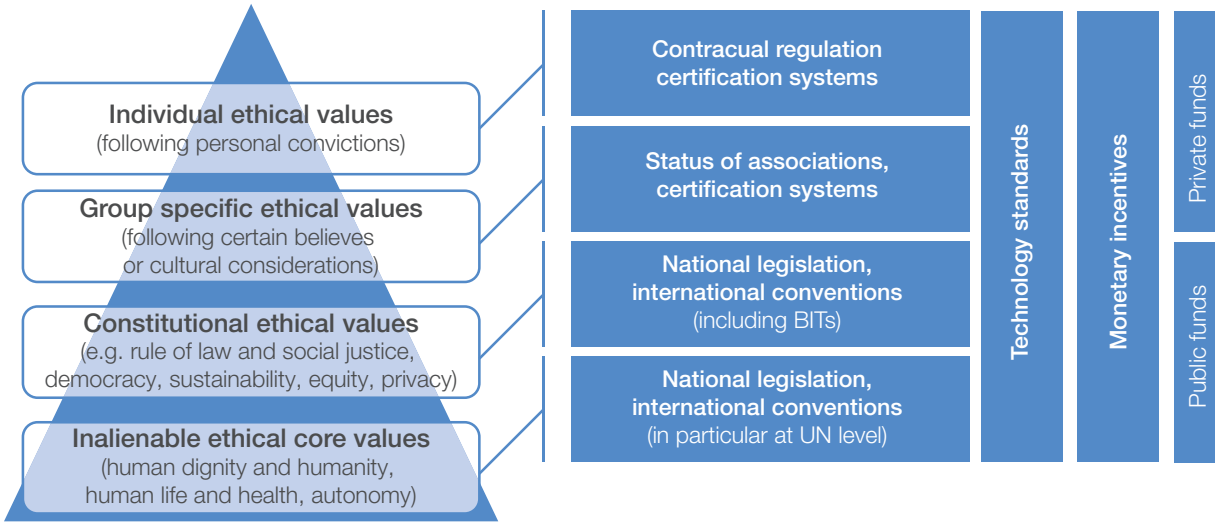
Good AI governance requires that the right regulatory instrument be chosen for each ethical concern. In view of the diversity of ethical values outlined above, it's clear there can be no "one-size-fits-all" solution. Formal legislation may in particular be required under such principles as the German constitutional principle of "*Parlamentsvorbehalt*" in case the use of new technologies has material implications on the protection of fundamental rights and constitutional principles.[93] Also, the obligation not to cause harm to other people, the need to compensate for damages in case harm is caused, as well as the obligation to respect personality rights and private autonomy and to protect privacy are generally subject to regulation by statutory laws on the national and international levels. In this regard, the precautionary principle may further call for binding legislative regulation.[94]

In contrast, individual ethical concerns following personal convictions might best be realized by individual, bilateral contractual agreements that are binding upon the parties to such agreement only. Value communities following group-specific convictions might be interested in the development of self-regulation-based certification systems that indicate compliance of products with relevant group-specific ethical values. For instance, whether an autonomous system was produced by sourcing sustainable resources and the exclusive use of renewable energy could be indicated by appropriate certificates. A further example is a smart home robot that could be programmed to only recommend suppliers of kosher food to its Jewish owners.

In addition to the various policy-making instruments explained above, the development of technological standards that allow technical solutions that comply with specific regulatory requirements should be considered. Adopting an AI design in care robots that respects the user's will as its guiding operational principle could be made by compliance with a relevant technology standard while a different standard could be developed for a more paternalistic AI system design. The kind of technology standard employed could be indicated to users by reference to a certain certificate. As already indicated, regulators should, in addition, consider granting specific monetary incentives to ensure the compliance of AI applications with ethical requirements. In particular, as AI is an emerging new technology, a particular adaptive approach could be to subject research and development funding grants to compliance with specific ethical principles. An example of rethinking legislation can be seen in the United Kingdom's Modern Slavery Act 2015. The government wanted to encourage truthful reporting by companies and thus enforced moderate rather than punitive fines.

Policy-makers should consider the diverse nature of ethical concerns and work on the basis of a graded governance system for ethical concerns in AI and autonomous systems to determine the appropriate content and technique for regulation. A graded governance model is illustrated in Figure 2:

Figure 2: Graded governance model for the implementation of ethical concerns in AI systems



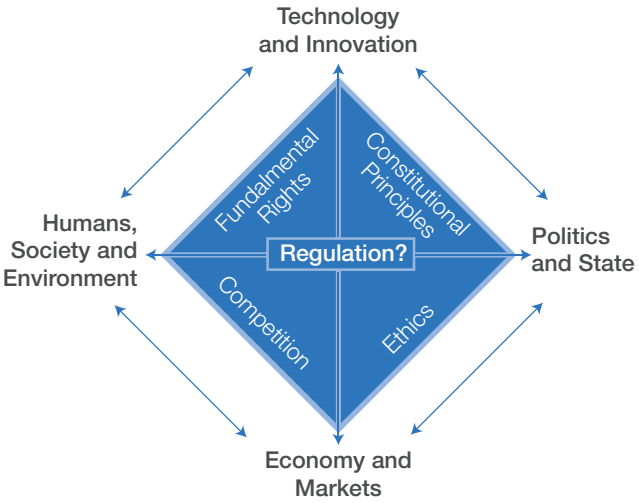
Source: Authors

c. The magic square of regulation in technological societies

AI governance is a particularly complex and difficult task also because all relevant parameters are directly or indirectly interrelated. The increasing use of AI and autonomous systems has a direct impact on humans, society and the environment. Existing jobs may become obsolete, new jobs will arise, less social interaction and more man-to-machine communication is expected and more raw materials may be consumed to increasingly build machines.[95] At the same time, new technologies bring about new business opportunities and can thus shape new markets or reshape existing ones. Depending on the nature of these new technologies' effects, governments may need to consider new regulatory actions. Regulation, however, implies a value decision must be made in light of various, sometimes even contradictory, fundamental principles. This includes the principle of competition, considered a key driver of consumer and public welfare, and further fundamental normative principles as expressed in basic rights, constitutional principles and ethics.

Particular difficulties arise because the actions or reactions of one of the aforementioned stakeholders can affect the other aspects and stakeholders. Also, as mentioned, regulation can have an impact on innovation dynamics. However, regulation may foster the development of new technologies and technology-focused business models. One example already referred to above is data privacy regulation, which on the one hand restricts the free use of personal data but on the other incentivizes businesses to develop privacy-by-design solutions, and thereby contributes to a high level of data protection. The relationship between the affected stakeholders and the principles to be referred to for regulation purposes can, therefore, best be described as a magic square, illustrated in Figure 3:

Figure 3: Magic square of regulation in technology-driven societies



Source: Authors

Finding the right regulatory solution within this magic square in view of new digital and AI-driven technologies is a particular challenge because the technology changes rapidly and no one knows at what stage it will be in five years. In addition, innovation cycles are generally extremely short in the field of digital technologies, including AI and autonomous systems, so the regulation itself is challenging in this field. Consequently, the governance mechanisms chosen must be agile.

3. The question of when to regulate

In view of the increasingly short innovation cycles, policy-makers must also deal with the question of when to regulate. Overhasty regulatory actions should be avoided, however. To efficiently and effectively protect fundamental rights and values, policy-makers need to ensure that the necessary regulation is implemented sufficiently early to avoid new technologies causing irreparable harm. One need only think of the hypothetical situation humanity would have met had there been forethought regarding the possible dangers associated with the use of nuclear energy. Having initially considered the potential abuses and possible ways to deal with nuclear waste would have avoided significant evils shadowing modern human history. This example illustrates that deliberating the possible dangers and methods in order to address and avoid them should be the first step when contemplating new technologies, in particular in cases involving AI whose operating modes and impacts cannot be entirely foreseen. It is now time to carefully evaluate the possible risks and ways to exclude or at least limit the risks. In particular, the precise definition of certain red lines should be considered for AI, as should whether, in view of the sensible application of the precautionary principle, AI algorithms, at least in certain use cases, should be subjected to an appropriate *ex ante* control system.[96]

Summary and outlook

The increasing use of AI and autonomous systems will have revolutionary effects on human society. Despite many benefits, AI and autonomous systems involve considerable risks that must be managed well to take advantage of their benefits while protecting ethical values as defined in fundamental rights and basic constitutional principles, thereby preserve a human-centric society. This White Paper advocates the need to conduct in depth risk–benefit assessments on the use of AI and autonomous systems and points out major concerns related to them, such as possible job losses, potential damages they might cause, a lack of transparency, the increasing loss of humanity in social relationships, the loss of privacy and personal autonomy, information biases, as well as error proneness and susceptibility to the manipulation of AI and autonomous systems. This analysis aims to raise policy-makers' awareness so they address these concerns and design an appropriate AI governance regime that preserves a human-centric society. Raising awareness of the eventual risks and concerns should not, however, be misunderstood as an anti-innovative approach. Rather, the risks and concerns must be considered adequately and sufficiently to make sure that such new technologies as AI and autonomous systems are built and operate in a way that is acceptable to individual human users and human society as a whole.

The variety of possible policy-making instruments underlines that ethical concerns are not necessarily best addressed by legislation or international conventions but that, depending on the ethical concern at hand, such alternative regulatory measures as technical standardization or certification may be preferable. For individual ethical concerns, bilateral contractual agreements may be sufficient. As suggested in this paper, an approach to develop a visionary AI governance regime could be to follow a graded governance model for the implementation of ethical concerns in AI systems. Good AI governance consists of a balanced policy mix with as much legislation as necessary and as much freedom as possible, combined with appropriate certification systems, technology standards and monetary incentives. Concerning the latter, regulators should in particular take their responsibility and messages seriously and only support research and development projects that comply with fundamental ethical principles and values.

Endnotes

1. Another question is whether and to what extent AI should be used for certain purposes, including to create automated weapon systems or humanoid robots, which requires an in-depth analysis and separate consideration.
2. Muehlhauser, L. and L. Helm, "Intelligence Explosion and Machine Ethics" (2012), Machine Intelligence Research Institute (MIRI), available at https://intelligence.org/files/IE-ME.pdf (accessed 16 August 2018), p. 2.
3. Moor, J., "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years", *AI Magazine*, vol. 27, no. 4 (2006), available at https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1911/1809 (accessed 16 August 2018).
4. McCarthy, J., M. Minsky, N. Rochester and C. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence", 31 August 1955, available at https://aaai.org/ojs/index.php/aimagazine/article/view/1904/ (accessed 11 October 2018).
5. European Commission, European Political Strategy Centre (EPSC) Strategic Notes, Issue 29, 27 March 2018, "The Age of Artificial Intelligence", p. 2.
6. UK Government Office for Science, *Artificial intelligence: opportunities and implications for the future of decision making* (2015), p. 5, available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf; also referred to in: UK Information Commissioner's Office, *Big data, artificial intelligence, machine learning and data protection* (2017), version 2.2, rec. 8, available at https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf (both accessed 16 August 2018).
7. UK Government Office for Science, op. cit. fn. 6.
8. UK Government Office for Science, op. cit. fn. 6, p. 6.
9. Castelvecchi, D., "Can we open the black box of AI", *Nature*, vol. 538, 6 October 2016, p. 22, available at http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731 (accessed 16 August 2018).
10. UK Government Office for Science, op. cit. fn. 6, p. 6.
11. UK Information Commissioner's Office, op. cit. fn. 6, rec. 16, p. 10.
12. UK Government Office for Science, op. cit. fn. 6, p. 7.
13. Tavani, H., *Ethics and Technology*, 5th Edition (2016), Wiley, p. 29; Sterba, J., *Ethics: the big questions*, p. 1.
14. Tavani, op. cit. fn. 13.
15. Brinkman, B. and A. Sanders, *Ethics in a Computing Culture* (2013), Cengage Learning, p. 7.
16. AI applications will be used in particular by people with disabilities and the elderly, in healthcare, agriculture and food supply, manufacturing, energy and critical infrastructure, logistics and transport, as well as in security and safety. European Parliamentary Research Service, Scientific Foresight Unit (STOA), "Ethical Aspects of Cyber-Physical Systems" (2016), p. 9. For the increasing relevance of AI applications, see European Commission, COM(2018)237 final, "Artificial Intelligence for Europe", no. 1.
17. European Parliamentary Research Service, op. cit. fn. 16, p. 14.
18. For an economic analysis, see Hildebrand, V., "Individualisierung als strategische Option der Marktbearbeitung – Determinanten und Erfolgswirkungen kundenindividueller Marketingkonzepte" (1997).
19. For details on this argument of supply-side substitutability, see Commission Notice on the definition of relevant market for the purposes of Community competition law (1997), OJ C 372/5, para. 20 et seqq.
20. With regard to challenges linked to the increasing use of computers, see Lin, P., K. Abney and G. Bekey, "Robot ethics: Mapping the issues for a mechanized world", *Artificial Intelligence*, vol. 175 (2011), p. 942.
21. Quinn, M., *Ethics for the Information Age*, 7th Edition (2016), p. 24, figure 1.12.
22. See European Parliament, resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), P8_TA(2017)005, Introduction, lit. I, J, K and L, available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN (accessed 16 August 2018).
23. German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung), *Zukunftsmonitor IV: Wissen schaffen – Denken und Arbeiten in der Welt von morgen*, published as of March 2017: 58% of a group of 1,004 participating German citizens believe that digitalization and robotics will cause more job losses than create new jobs (p. 3); 80% believe that the main tasks of routine jobs will be performed by machines or computer programs in the year 2030 (p. 4); and 81% expect that due to the speed of technological change, more and more people will become increasingly isolated (p. 6).
24. McKinsey&Company, *A Future that works: Automation, Employment, and Productivity* (2017), p. 5.
25. Berriman, R. and J. Hawksworth, PwC, "Will robots steal our jobs? The potential impact of automation on the UK and other major economies" (2017), p. 1. The authors state that "up to 30% of UK jobs could potentially be at high risk of automation by the early 2030s", while various figures apply for other economies (US: 38%, Germany: 35%, Japan: 21%).
26. PwC, *UK Economic Outlook* (2018), no. 4.8, p. 49, available at https://www.pwc.co.uk/economic-services/ukeo/ukeo-july18-full-report.pdf (accessed 16 August 2018).
27. European Commission, SWD(2018) 137 final, "Liability for emerging digital technologies", no. 1; European Commission, (2018) 237 final, "Artificial Intelligence for Europe", no. 3.3 ("Safety and liability"); European Parliamentary Research Service, op. cit. fn. 16, p. 8.
28. For European law, see in particular the Council Directive of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, OJ 1985, L 210/29 and Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, OJ 2006, L 157/24 as the relevant European safety legislation in relation to robots. For further relevant legislation, see European Commission, "Liability for emerging digital technologies", SWD(2018) 137 final, no. 2.1.

29. According to a 2017 European Commission consultation, GROW/B1/HI/sv(2017) 3054035, "45% of producers, 58% of consumers and 44% of the other respondents (including public authorities and civil society) consider that for some products (e.g. products where software and applications from different sources can be installed after purchase, products performing automated tasks based on algorithms, data analytics, self-learning algorithms or products purchased as a bundle with related services), the application of the Directive might be problematic or uncertain." For an analysis, see European Commission, SWD(2018) 137 final, in particular nos. 2 and 4.
30. IEEE, *Ethically Aligned Design – A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, version 2 (2017), p. 136 et seqq., available at http://standards.ieee.org/develop/indconn/ec/ead_v2.pdf, also downloadable via https://ethicsinaction.ieee.org/ (accessed 16 August 2018), p. 148 et seqq.; European Parliament, op. cit. fn. 22, rec. 49 et seqq.
31. See, for example, Bostrom, N. and E. Yudkowsky, "The Ethics of Artificial Intelligence", p. 1, available at https://intelligence.org/files/EthicsofAI.pdf (accessed 16 August 2018). The lack of transparency is in particular due to the technical design of deep learning mechanisms; see above section I.1.a).
32. UK Information Commissioner's Office, op. cit. fn. 6, rec. 16. For issues related to the black box effect in AI algorithms used for medicinal purposes, see Price II, W., "Black Box Medicine", *Harvard Journal of Law & Technology*, vol. 28, no. 2 (2015), p. 419, 432.
33. For the reasons, see the description of the purpose for developing the IEEE P7001™ standard, no. 5.4, Approved PAR, available online at https://standards.ieee.org/develop/project/7001.html (accessed 16 August 2018).
34. Bostrom and Yudkowsky, op. cit. fn. 31, p. 1 et seq.
35. Cathelat, B., in "Human Decisions – Thoughts on AI", UNESCO Publishing (2018), available at http://unesdoc.unesco.org/images/0026/002615/261563e.pdf (accessed 16 August 2018), p. 132, 134.
36. Lin, P., K. Abney and G. Bekey, *Artificial Intelligence*, vol. 175 (2011), p. 942, no. 3.3.
37. Groth, O., M. Nitzberg and M. Esposito, "Rules for Robots", p. 18, available at http://www.kas.de/wf/doc/kas_52115-544-2-30.pdf?180418140416 (accessed 16 August 2018). The authors state: "The cognitive ability of AI will transform human existence over the next 10 to 20 years."
38. European Parliamentary Research Service, op. cit. fn. 16, p. 8.
39. Drexl, J. et al., "Position Statement of the Max Planck Institute for Innovation and Competition of 26 April 2017 on the European Commission's 'Public consultation on Building the European Data Economy'", p. 2, 9 et seqq., available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2959924 (accessed 16 August 2018); European Commission, COM(2018)237 final, "Artificial Intelligence for Europe", no. 3.1 ("Making more data available").
40. European Parliament, op. cit. fn. 22, rec. 20 et seqq.; The White House, Washington, Executive Office of the President, "Big Data: Seizing Opportunities, Preserving Values" (2014), p. 61; "Written Testimony of Frank Pasquale, University of Maryland, Before the United States House of Representatives, Committee on Energy and Commerce, Subcommittee on Digital Commerce and Consumer Protection, 'Algorithms: How Companies' Decisions About Data and Content Impact Consumers' 29 November 2017, p. 5 et seqq.
41. European Commission, COM(2018)237 final, op. cit. (fn. 39), no. 3.1 ("Making more data available").
42. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016, L 119/1.
43. United States District Court Northern District of California, Cahen v. Toyota Motor Corp., 3:15-cv-01104 (N.D. Cal. March 10, 2015); United States District Court for the Southern District of Illinois, Flynn v. FCA US LLC., 3:15-cv-855 (S.D. Ill. 4 August 2015).
44. Spitzer, M., *Digitale Demenz: Wie swir uns und unsere Kinder um den Verstand bringen* (2012); Dossey, L., *EXPLORE* (March/April 2014), vol. 10, no. 2, p. 69; less critical: Appel, M. and C. Schreiner, *Psychologische Rundschau* (2014), 65, p. 1 et seqq.
45. See, for example, the iPhone app Koubachi, available at https://itunes.apple.com/de/app/koubachi-pers%C3%B6nlicher-pflanzenpflege-assistent/id391581160?mt=8 (accessed 16 August 2018).
46. Daniel, M. and D. Striebel, *Künstliche Intelligenz, Expertensysteme – Anwendungsfelder, neue Dienste, soziale Folgen* (1993), p. 103.
47. European Commission, COM(2018)237 final, "Artificial Intelligence for Europe", nos. 1 and 3 (see also fn. 51 of the European Commission document); Stucke, M. and A. Ezrachi, "How Digital Assistants Can Harm Our Economy, Privacy And Democracy", p. 1271, *Berkeley Technology Law Journal* (2017), vol. 32, no. 3, 1239, 1271; Bostrom and Yudkowsky, op. cit. fn. 31; see also Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter and L. Floridi, "The Ethics of Algorithms: Mapping the Debate", *Big Data & Society* (2016), available at http://journals.sagepub.com/doi/pdf/10.1177/2053951716679679 (accessed 16 August 2018), p.7, in which the authors state that against this background "[a]lgorithms inevitably make biased decisions."
48. EPSC Strategic Notes, op. cit. fn. 5, p. 7.
49. Holznagel, B., "Phänomen 'Fake News': Was ist zu tun? Ausmaβ und Durchschlagskraft von Desinformationskampagnen", *MultiMedia und Recht* (2018), 18, 19; Paal, B. and M. Hennemann, "Meinungsvielfalt im Internet - Regulierungsoptionen in Ansehung von Algorithmen, Filterblasen, Fake News und Social Bots", *Zeitschrift für Rechtspolitik* (2017), 76, 77; for a detailed analysis, see Drexl, J., "Neue Regeln für die Europäische Datenwirtschaft", *NZKart - Neue Zeitschrift für Kartellrecht* (2017), 339 et seqq., 415 et seqq.
50. Ambrose, M. L., "The Law and the Loop" (2014), Communication, Culture & Technology, Georgetown University, II.B. The issue also becomes obvious from the current discussions around possible manipulations of the US presidential election 2016; see, for instance, Todd, C. and C. Dann, "How Big Data Broke American Politics", NBC News, 14 March 2017, available at https://www.nbcnews.com/politics/elections/how-big-data-broke-american-politics-n732901 (accessed 16 August 2018) and Stucke and Ezrachi, op. cit. fn. 47, p. 1273.
51. For details, see OECD, "Algorithms and Collusion – Background note by the Secretariat", OECD Directorate for Financial and Enterprise Affairs Competition Committee, 21-23 June 2017, available at https://one.oecd.org/document/DAF/COMP(2017)4/en/pdf (accessed 16 August 2018).
52. Quinn lists typical examples in *Ethics for the Information Age*, op. cit. fn.21, p. 365 et seqq.; also see Bostrom and Yudkowsky, op. cit. fn. 31, p. 2.
53. For the definition of AI, see above no. I.1.a.

54. Hofstetter, Y., *Das Ende der Demokratie: Wie die künstliche Intelligenz die Politik übernimmt und uns entmündigt* (2016), p. 361. The author states: "Die Einschätzung der Künstlichen Intelligenz wird dabei nicht immer zutreffen. Sie nehmen eine generelle Klassifizierung menschlichen Verhaltens vor, die auf Statistik beruht und deshalb von Unschärfe, das heißt Fehleinschätzungen, betroffen ist. ["The assessments made by AI will not always be correct. It classifies human behaviour on the basis of statistics and thus the classification is affected by uncertainty, that is, misjudgements."]"
55. The German Federal Supreme Court stated expressly that extrapolating from statistical data to individual cases causes general difficulties and that it is usually impossible to decide on the basis of statistical data whether the result of a specific assessment is correct; Federal Court of Justice (*Bundesgerichtshof*), decision of 17 December 1998, NJW 1999, 657, 658, 661.
56. See Bostrom and Yudkowsky, op. cit. fn. 31, p. 2; Stanford University, "One Hundred Year Study on Artificial Intelligence (AI100)", available at https://ai100.stanford.edu/ (accessed 16 August 2018), p. 42. For examples, see German Federal Office for Information Security, *The State of IT Security in Germany 2017*, p. 14 (regarding the possible manipulation of traffic lights) and p. 16 (regarding the possible manipulation of smart home components for access control, which may be attacked and manipulated to prepare a burglary), available at https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Securitysituation/IT-Security-Situation-in-Germany-2017.pdf?__blob=publicationFile&v=3 (accessed 16 August 2018).
57. Zeit online, "Ein Freies und sicheres Web für alle [A Free and Secure Web for Everyone]", available at http://www.zeit.de/digital/datenschutz/2018-02/hacker-dringen-in-deutsches-regierungsnetz-ein (accessed 16 August 2018).
58. German Federal Office for Information Security, op. cit. (fn. 56), p. 75.
59. Polonski, V., "Artificial intelligence can save democracy unless it destroys it first", Oxford Internet Institute blog comment, 10 August 2017, available at https://www.oii.ox.ac.uk/blog/artificial-intelligence-can-save-democracy-unless-it-destroys-it-first/ (accessed 16 August 2018).
60. Smith, B., "Facial recognition technology: The need for public regulation and corporate responsibility", Microsoft blog comment, 13 July 2018, available at https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/ (accessed 16 August 2018).
61. OECD, "Algorithms and Collusion: Competition Policy in the Digital Age" (2017), available at http://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm, in particular p. 18 et seqq. (accessed 16 August 2018).
62. Etzioni, A. and O. Etzioni, "Incorporating Ethics into Artificial Intelligence", *The Journal of Ethics*, vol. 17, no. 3 (2017), DO 10.1007/s10892-017-9252-2, no. 1.2., available at http://ai2-website.s3.amazonaws.com/publications/etzioni-ethics-into-ai.pdf (accessed 21 October 2018); Allen, C., I. Smit and W. Wallach, "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches", *Ethics and Information Technology*, vol. 7, no. 3 (2005), pp. 149, 150.
63. Etzioni and Etzioni, op. cit. fn. 62.
64. The Guardian, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", 24 March 2016, available at https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (accessed 16 August 2018).
65. Etzioni and Etzioni, op. cit. fn. 62.
66. United Nations, "Universal Declaration of Human Rights", Art. 3, available at http://www.un.org/en/universal-declaration-human-rights/ (accessed 16 August 2018); "Charter of Fundamental Rights of the European Union", Art. 2, OJ 2000, C 364/1, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (accessed 16 August 2018).
67. Neuhäuser, C., "Roboter und moralische Verantwortung" in: Eric Hilgendorf (ed.), *Robotik im Kontext von Recht und Moral*, p. 281.
68. "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 to 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." Quoted from Weaver, J. F., *Robots are People Too* (2014), p. 4.
69. See, for instance, Bible, English Standard Version (ESV), Matthew 7:12: "So whatever you wish that others would do to you, do also to them, for this is the Law and the Prophets."; for an overview, see Lepard, B., *Hope for a Global Ethic* (2005), p. 35 et seqq. and "Die 'Goldene Regel' in den Weltreligionen", available at https://www.erzdioezese-wien.at/dl/OKrlJKJlMnklJqx4kJK/11JKW_Goldene_Regel_Zivilcourage_konkret_-_Schulmodul.pdf (accessed 16 August 2018).
70. Allen, C., W. Wallach and I. Smit, "Why Machine Ethics?", IEEE Intelligent Systems, vol. 21, no. 4, July-August 2006, p. 12, 14; Mittelstadt, Allo, Taddeo, Wachter and Floridi, op. cit. fn. 47, p. 12. The authors underline that no consensus exists on how to practically relocate the social and ethical duties displaced by automation.
71. Neuhäuser, op. cit. fn. 67, p. 282.
72. Mittelstadt, Allo, Taddeo, Wachter and Floridi, op. cit. fn. 47, p. 12.
73. Mittelstadt, Allo, Taddeo, Wachter and Floridi, op. cit. fn. 47, p. 13; Cath, C. et al., "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach" (23 December 2016), available at SSRN: https://ssrn.com/abstract=2906249, p. 2, lit. (a) (accessed 16 August 2018), p. 20.
74. Scherer, M., "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies", *Harvard Journal of Law & Technology*, vol. 29, no. 2, spring 2016, p. 395.
75. Etzioni and Etzioni appear to take a different view when they state that "there is little need to teach machines ethics even if this could be done in the first place", op. cit. fn. 62, Abstract. This view, however, is not convincing as AI system compliance with ethical requirements requires the technical implementation of the corresponding requirements, all the more when machines act increasingly without direct human control.
76. Particularities have to be taken into account in view of international political entities, such as the European Union, where national Member States have referred selected sovereign powers to the European Union as an international body having the (limited) competence to enact laws that are automatically binding within all Member States, see Art. 288 TFEU.
77. EPSC Strategic Notes, op. cit. fn. 5, p. 6, with regard to the data protection standards as established by the EU General Data Protection Regulation, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, OJ 2016, L 119/1. As a side note, this was confirmed by a number of selected industry companies (methodologically based on qualitative interviews conducted with selected industry representatives in the research group on data driven markets of the Max Planck Institute for Innovation and Competition, Munich). According to this, companies took the view that the strict European privacy regulation can

amount to a potential advantage in international competition. The reason is that business models complying with European data protection rules may be more acceptable for consumers. Therefore, in particular in combination with corresponding certificates, strict regulation can – certainly depending on the circumstances of each case – foster economic growth and thereby public and private wealth.

78. An example of an international initiative aimed at a new governance regime for AI-driven systems is the recent EU Parliament's initiative on civil law rules for robots: European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), P8_TA(2017)005; European Commission, COM(2018)237 final; the UN has established its "Centre for Artificial Intelligence and Robotics" in The Hague which will, among other things, perform a risk assessment and stakeholder mapping and analysis: see UNICRI Centre for Artificial Intelligence and Robotics, available at http://www.unicri.it/in_focus/on/UNICRI_Centre_Artificial_Robotics (accessed 16 August 2018); specifically, there is an ongoing debate around an international ban of AI-driven killer robots: Weaver, *Robots are People Too* (2014), p. 142 et seqq.; Walsh, T., "Why the United Nations Must Move Forward With a Killer Robots Ban", 15 December 2016, available at http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/united-nations-killer-robots-ban (accessed 16 August 2018).
79. It could, for instance, be agreed upon therein that certain protective measures will be established in relation to AI systems, such as marketing authorization requirements for certain AI systems, requirements to provide for certain strict liability regimes to recover damages caused by AI systems or requirements to provide for transparency as regards the functioning and decision-making processes used by AI systems.
80. With regard to AI and autonomous systems, technology standards could be developed that make use of technical measures providing for ethical compliant behaviour by AI algorithms. That includes privacy by design or transparency by design solutions as well as potential kill switch technologies.
81. As an alternative to binding legislative measures, public international organizations can create soft law, such as guidelines on ethical compliance of AI systems. A major advantage is that other than binding and enforceable statutory rules, guidelines and similar soft law may be established in less complex procedures.
82. See "Agile Governance: Reimagining Policy-making in the Fourth Industrial Revolution", World Economic Forum White Paper, available at https://www.weforum.org/whitepapers/agile-governance-reimagining-policy-making-in-the-fourth-industrial-revolution (accessed 16 August 2018).
83. With regard to AI applications, regulators could, for example, subject the grant of research and development funding to the compliance of respective research and development proposals and their results with specific ethical requirements. To this end, the relevant core ethical principles would need to be defined in a first step, for instance within the framework of an ethics charter for AI applications; see, in particular, the work done by the European Group on Ethics in Science and New Technologies, below no. IV.2.a. According to the European Commission, COM(2018)237 final, no. 3.3, "AI ethics guidelines" should be developed by the end of 2018.
84. Chatila, R. et al., "IEEE Global Initiative Aims to Advance Ethical Design of AI and Autonomous Systems", 29 March 2017, available at https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/ieee-global-initiative-ethical-design-ai-and-autonomous-systems (accessed 16 August 2018).
85. IEEE, op. cit. fn. 30; for Version 1 of the document *Ethically Aligned Design* (2016), see http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf (accessed 16 August 2018).
86. For an overview of the standardization projects of the P7000 series, see IEEE, op. cit. fn. 30, p. 4, and Ethics in Action, available at https://ethicsinaction.ieee.org/ (accessed 16 August 2018).
87. IEEE, op cit. fn. 30, p. 3.
88. Rozenfeld, M., "Seven IEEE Standards Projects Provide Ethical Guidance for New Technologies", 5 May 2017, available at http://theinstitute.ieee.org/resources/standards/seven-ieee-standards-projects-provide-ethical-guidance-for-new-technologies (accessed 16 August 2018).
89. For the need to conduct a risk–benefit assessment, see also von Schomberg, R., "From the Ethics of Technology towards an Ethics of Knowledge Policy & Knowledge Assessment" (2007), p. 15 et seqq., available at https://publications.europa.eu/en/publication-detail/-/publication/aa44eb61-5be2-43d6-b528-07688fb5bd5a (accessed 16 August 2018).
90. For me details on these ethical concerns, see European Group on Ethics in Science and New Technologies (EGE), "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems" (2018), p. 5, available at http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf (accessed 16 August 2018).
91. "Universal Declaration of Human Rights", op. cit. fn. 66; "Convention on the Rights of the Child, Adopted and opened for signature, ratification and accession by General Assembly resolution 44/25 of 20 November 1989", available at http://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_44_25.pdf (accessed 16 August 2018).
92. German Federal Network Agency, "Bundesnetzagentur removes children's doll 'Cayla' from the market", decision of 17 February 2017, press release available at https://www.bundesnetzagentur.de/SharedDocs/Pressemitteilungen/EN/2017/17022017_cayla.html (accessed 16 August 2018).
93. Grundgesetz [GG] [Basic Law], Arts. 1–20, 79(3), see Jarass, H. and B. Pieroth, *Grundgesetz für die Bundesrepublik Deutschland: GG*, German Constitution (13th Ed. 2014); von Westphalen, R., *Parlamentslehre*, vol. 2 (1996), p. 433 et seqq., p. 439.
94. European Commission, Communication on the Precautionary Principle, COM(2000)1 final, p. 8 et seqq. (with regard to Europe) and p. 10 et seqq. (with regard to the WTO), p. 13 et seqq. These texts underline the precautionary principle as a basic rule that aims to protect consumers against potential harmful developments on the basis of scientific risk assessments; the importance of the precautionary principle is also underlined by the Court of Justice of the European Union, decision of 5 May 1998, C-157/96 and C-180/96, rec. 63 resp. rec. 99 – *BSE*; Neuhäuser, op. cit. fn. 67, p. 284; Weckert, J., "In Defense of the Precautionary Principle", *IEEE Technology and Society Magazine* (2012), p. 12 et seqq., available at https://researchoutput.csu.edu.au/ws/portalfiles/portal/8858395 (accessed 16 August 2018). It should be noted, though, that the precautionary principle is still not entirely acknowledged as a governance principle in international law; Bourguignon, D., European Parliamentary Research Service (EPRS), "The precautionary principle: Definitions, applications and governance" (2015), p. 6.
95. For the various concerns associated with the increasing use of AI and autonomous systems, see section I.3.
96. Note in particular the suggestion made by Smith, op. cit. fn. 60.

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744

contact@weforum.org
www.weforum.org