ARTICLE: IMPLEMENTING ETHICS INTO ARTIFICIAL INTELLIGENCE: A CONTRIBUTION, FROM A LEGAL PERSPECTIVE, TO THE DEVELOPMENT OF AN AI GOVERNANCE REGIME

December 20, 2019

Reporter
18 Duke L. & Tech. Rev. 176 *

Length: 46857 words

Author: DR. AXEL WALZ + & KAY FIRTH-BUTTERFIELD ++

+ District Attorney, Public Prosecutor's Office Traunstein, former Senior Research Fellow, Max Planck Institute for Innovation and Competition, Munich

++ Head of Artificial Intelligence and ***Machine*** ***Learning***, World Economic Forum, San Francisco

Highlight

ABSTRACT

*The increasing **use** of AI and autonomous systems will have revolutionary impacts on society. Despite many benefits, AI and autonomous systems involve considerable risks that need to be managed. Minimizing these risks will emphasize the respective benefits while at the same time protecting the **ethical** values defined by fundamental rights and basic constitutional principles, thereby preserving a human centric society. This Article advocates for the need to conduct in-depth risk-benefit-assessments with regard to the **use** of AI and autonomous systems. This Article points out major concerns in relation to AI and autonomous systems such as likely job losses, causation of damages, lack of transparency, increasing loss of humanity in social relationships, loss of privacy and personal autonomy, potential information biases and the error proneness, and susceptibility to manipulation of AI and autonomous systems. This critical analysis aims to raise awareness on the side of policy-makers to sufficiently address these concerns and design an appropriate AI governance regime with a focus on the preservation of a human-centric society. Raising awareness for eventual risks and concerns should, however, not be misunderstood as an anti-innovative approach. Rather, it is necessary to consider risks and concerns adequately and sufficiently in order to make sure that new technologies such as AI and autonomous systems are constructed and operate in a way which is acceptable for individual users and society as a whole. To this end, this article develops a graded governance model for the implementation of **ethical** concerns in AI systems reflecting the often-misjudged fact that, actually, there is a variety of policy-making instruments which policy-makers can make **use** of. In particular, **ethical** concerns do not only need to be addressed by legislation or international conventions. Depending on the **ethical** concern at hand, alternative regulatory measures such as technical standardization or certification may even be preferable. To illustrate the practical impact of this graded governance model for the implementation of **ethical** concerns in AI systems, two concrete global approaches are presented herein, in addition, which regulators, governments and industry could refer to as a basis for regulating **ethical** concerns associated with the **use** of AI and autonomous systems*.

Text

[*180] INTRODUCTION

The relationship between mankind and ***machines*** has been a subject of emotional debates and visionary utopian poetry for many centuries, full of hopeful fascination and apocalyptic anxiety. [1]The discussion certainly became more intense with the industrialization in the 19th and early 20th centuries, [2]and it is becoming more urgent as a consequence of ever more digitalization and the implementation of artificial intelligence ("AI"). [3]Therefore, we are not looking at an entirely new debate when we ask ourselves--often slightly critically and skeptically--what role we may have to or may be able to play when ***machines*** and automated systems take over more and more tasks originally performed by us. [4]This discussion is probably more relevant than ever in view of the intensity of the expected automation on the basis of the implementation of AI-driven technologies. [5]Driven by stronger computational power, more sophisticated **[*181]** algorithms and higher availability of vast amounts of good quality data, ***machines*** are increasingly able to act independently without human command. [6]Moreover, AI-driven systems act on the basis of self-***learning*** algorithms that enable them to perform in increasingly autonomous and often unexpected ways. This may enable AI to ultimately make, or at least influence decisions, that may conflict with our general ***ethical*** principles and values. [7]***Ethical*** principles developed over centuries of history through difficult efforts despite strong resistance from the ruling class. It is an axiomatic assumption that irrespective of digitalization and automation, these ***ethical*** principles and values shall be preserved. Likewise, we assume there is a common understanding that new technologies should be ***used*** to further promote and establish ***ethical*** values and principles as basic guidelines for our daily life and be ***used*** to thereby develop "a good AI society." [8]

---

[1] Popular characters to be referred to in this regard are "Golem," "Frankenstein" and more recent works such as "Terminator," "Transformers" and "I, Robot." For a comprehensive overview of the literary and artistic discussion of the relationship between humans and ***machines***, see ULRIKE BARTHEMEß & ULRICH FUHRBACH, IROBOT - UMAN: KÜNSTLICHE INTELLIGENZ UND KULTUR: EINE JAHRTAUSENDEALTE BEZIEHUNGSKISTE (2012).

[2] An interesting artistic examination of the increasing degree of automation and industrialization of manufacturing processes is Charlie Chaplin's film MODERN TIMES (Charles Chaplin Productions 1936). For a more detailed description, see *Modern Times*, INTERNET MOVIE DATABASE, http://www.imdb.com/title/tt0027977/ (last visited Oct. 6, 2018).

[3] Marshal S. Willick, *Artificial Intelligence: Some Legal Approaches and Implications*, AI MAGAZINE, Summer 1983, at 5. *See also* YVONNE HOFSTETTER, DAS ENDE DER DEMOKRATIE: WIE DIE KÜNSTLICHE INTELLIGENZ DIE POLITIK ÜBERNIMMT UND UNS ENTMÜNDIGT (2016).

[4] WINFRIED OPPELT, MENSCH, AUTOMAT UND AUTOMATISIERUNG, IN: MÖGLICHKEITEN UND GRENZEN DER AUTOMATION 31 (1965), already stated: " *Außerdem muss die Frage studiert werden, ob nicht durch die Automation noch viel tiefgreifende Wandlungen und Wirkungen ausgelöst werden, die den Standort des Menschen innerhalb der Schöpfung grundlegend verändern, z. B. durch die sogenannten denkenden Maschinen, zwangsläufige Entwicklungen, die kaum noch oder nicht mehr gesteuert werden können*."

[5] *See* EUROPEAN PARLIAMENTARY RESEARCH SERV.: SCI. FORESIGHT UNIT, ***ETHICAL*** ASPECTS OF CYBER-PHYSICAL SYSTEMS 36 (2016) [hereinafter EPRS]; Brent D. Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3 BIG DATA & SOCIETY 1 (2016), http://journals.sagepub.com/doi/pdf/10.1177/2053951716679679; Boer Deng,***Machine*** *Ethics: The Robot's Dilemma*, NATURE (July 1, 2015), https://www.nature.com/news/***machine***-ethics-the-robot-s-dilemma-1.17881.

[6] EUROPEAN POLITICAL STRATEGY CTR., THE AGE OF ARTIFICIAL INTELLIGENCE 1 (2018) [hereinafter EPSC].

[7] SETH BAUM, SOCIAL CHOICE ETHICS IN ARTIFICIAL INTELLIGENCE 1 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046725.

In a declaration on April 10, 2018, 25 EU Member States expressed their will to ensure "an adequate legal and ***ethical*** framework, building on EU fundamental rights and values" and to ensure that "humans remain at the centre of the development, deployment and decision-making of AI." [9]The European Commission's Group on Ethics in Science and New Technologies pointed out that this requires a "collective, wide-ranging and inclusive process of reflection and dialogue" focusing "on the values around which we want to organize society and on the role that technologies should play in it." [10]This Article hopes to enrich this debate by looking at possible means and mechanisms for implementing ***ethical*** values in AI-driven technology in order to contribute to building a human-centric AI-society. [11]The goal is to outline approaches on how to determine an AI governance regime that fosters the benefits of AI yet considers the relevant risks arising from the ***use*** of AI and autonomous systems. To this end, this Article **[*182]** posits different concepts that could be applied to ensure that the ***use*** of AI does not conflict with ***ethical*** values. The first section of this Article will illustrate certain ***ethical*** concerns regarding the ***use*** of AI. The second section will outline and discuss the advantages and downsides of different governance instruments that could be referred to in order to implement ethics in AI applications. The third section will present various practical approaches for governance of AI applications. Based on these insights, the fourth section concludes with recommendations as to how a holistic AI governance regime could be developed.

I. DEFINITION OF BASIC TERMS

It is necessary to define some basic terms before explaining potential benefits and risks in AI applications.

A. Definition of AI

While an intense discussion is ongoing about the possible regulation of AI, there is still a lack of widespread agreement on the definition of AI. [12]AI as a term was first coined by John McCarthy in the Dartmouth Summer Research Project of 1956. [13]McCarthy defined AI as a ***machine*** that behaves "in ways that would be called intelligent if a human were so behaving." [14]This definition, however, does not say anything about the technical functionality of AI. Focusing more on a technology's ability to adapt to changing circumstances, a further definition of AI refers to a "technology (software, algorithm, a set of processes, a robot, etc.) that is able to function appropriately with foresight of its environment." [15]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

---

[8] CORINNE CATH ET AL., ARTIFICIAL INTELLIGENCE AND THE 'GOOD SOCIETY': THE US, EU, AND UK APPROACH 2 (2016), https://ssrn.com/abstract=2906249.

[9] *Declaration of Cooperation on Artificial Intelligence*, EUROPEAN COMMISSION (Apr. 10, 2018), https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence.

[10] EUROPEAN GRP. ON ETHICS IN SCI. AND NEW TECH., STATEMENT ON ARTIFICIAL INTELLIGENCE, ROBOTICS AND "AUTONOMOUS SYSTEMS" (2018), http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

[11] A different question is whether and to what extent AI shall be ***used*** at all for certain purposes. This question, relating, e.g., to the admissibility of ***using*** AI in automated weapon systems or creating humanoid robots, requires in-depth analysis and needs to be dealt with separately.

[12] *See* LOUIE HELM & LUKE MUEHLHAUSER, INTELLIGENCE EXPLOSION AND ***MACHINE*** ETHICS 2 (2012), https://intelligence.org/files-IE-ME.pdf.

[13] James Moor, *The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years, in* 27 AI MAG. 87, 87 (2006), https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1911/1809.

[14] J. MCCARTHY ET AL., A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE 11 (1955), http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

[15] EPSC, *supra* note 6, at 2.

and anticipate possible future events." [16]This **[*183]** involves the creation of statistical models that are ***using*** series of algorithms or step-by-step instructions which computers can follow to perform a particular task. [17]

Technically, AI is mainly powered by ***machine learning*** algorithms, i.e., algorithms that change in response to their own received inputs and consequently improve with experience. [18]***Machine learning*** must 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, and 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. [19]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. [20]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 anymore (the so-called black box effect). [21]The multi-layered approach allows corresponding ***machines*** to not only follow pre-programmed decisions but also to respond to changes within their environment. Examples of this technology include the facial recognition systems referred to above and autonomous cars, which can make real-time decisions about speed and direction by administering sensor-based data without input from a human user. [22]

In summary, 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 produces specific output independent of human control.

[*184] *B. Definition of Ethics*

Ethics is commonly referred to as the study of morality. [23]Morality, as ***used*** in this Article, is a system of rules and values for guiding human conduct, as well as principles for evaluating those rules. [24]Consequently, ***ethical*** behavior does not necessarily mean "good" behavior. ***Ethical*** behavior instead indicates compliance with

---

[16] GOV'T OFFICE FOR SCI. (UK), ARTIFICIAL INTELLIGENCE: OPPORTUNITIES AND IMPLICATIONS FOR THE FUTURE OF DECISION MAKING 5 (2016), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf; INFO. COMM'R'S OFFICE, BIG DATA, ARTIFICIAL INTELLIGENCE, ***MACHINE LEARNING*** AND DATA PROTECTION (2017), https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

[17] GOV'T OFFICE FOR SCI., *supra* note 16, at 5.

[18] *Id*. at 5-6.

[19] David Castelvecchi, *Can We Open the Black Box of AI*?, NATURE, (Oct. 5, 2016), http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[20] GOV'T OFFICE FOR SCI., *supra* note 16, at 6.

[21] INFO. COMM'R'S OFFICE, *supra* note 16, at 10.

[22] GOV'T OFFICE FOR SCI., *supra* note 16, at 7.

[23] *See* HERMAN TAVANI, ETHICS AND TECHNOLOGY: ***ETHICAL*** ISSUES IN AN AGE OF INFORMATION TECHNOLOGY AND COMMUNICATION TECHNOLOGY 3 (2004); JAMES STERBA, ETHICS: THE BIG QUESTIONS 1 (1998).

[24] *See* TAVANI, *supra* note 23, at 35.

specific values. [25]Such values can be 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's specific code of conduct). This broad definition is ***used*** here because this Article does not approach AI from a specific normative perspective and does not analyze AI in a moral sense; rather, this Article seeks to contribute to the discussion around the determination of appropriate regulatory means in order to implement ethics into AI. In addition, the benefit of this neutral definition of ethics is that it enables one to address the issue of ***ethical*** diversity from a regulatory and policymaking perspective.

II. ***ETHICAL*** CONCERNS IN AI APPLICATIONS

A. Potential Benefits of AI Applications

A recent study, conducted on behalf of the European Parliament, concludes that AI applications will be ***used*** in almost all fields of our daily lives. [26]In each field, AI can provide benefits, including the reduction of economic inefficiencies and labor costs as well as an increase in high-skilled jobs. Moreover, AI can help companies understand their customers better and accordingly develop more customized products tailored to the specific needs of individual customers. The increasing flexibility of smart factories is likely to **[*185]** play a decisive role in this regard. [27]Better understanding of individual consumer needs allows for the development of more economically efficient sales and marketing strategies. [28]

While these benefits appear to favor the company side of modern economic systems, AI applications may also provide benefits to consumers. These benefits may predominantly depend on where, and how, AI is to be applied. By way of example, looking at the individualization of the manufacturing process, one benefit to consumers is the increase in the variety of products. The flexibility associated with the implementation of smart factories further increases competition between companies that might previously not have been considered as competitors. [29]Increasing competition can ultimately force companies to pass on an AI-driven reduction of production costs to their customers and result in lower prices.

B. Potential Risks of AI Applications

Despite the various potential benefits, AI poses a number of serious risks. These risks must be explored to ensure that human values can be sufficiently protected. Given AI's possible disruptive impacts, society will only trust and ***use*** AI subject to appropriate means of protection. [30]The risks, as well as the potential benefits, of AI applications strongly depend on the particular case. Still, several common areas of concern exist, which are summarized below.

1. Loss of Jobs

---

[25] WILLIAM J. BRINKMAN & ALTON F. SANDERS, ETHICS IN COMPUTING CULTURE 7 (2013).

[26] Including applications for disabled people and the daily life of elderly people, healthcare, agriculture and food supply, manufacturing, energy and critical infrastructure, logistics and transport as well as security and safety. EPRS, *supra* note 5, at 9. For more information concerning the increasing relevance of AI applications, see Commission Communication on Artificial Intelligence for Europe, COM (2018) 237 final (Apr. 25, 2018) [hereinafter Artificial Intelligence for Europe].

[27] EPRS, *supra* note 5, at 14.

[28] For an economic analysis, see VOLKER G. HILDEBRAND, INDIVIDUALISIERUNG ALS STRATEGISCHE OPTION DER MARKTBEARBEITUNG: DETERMINANTEN UND ERFOLGSWIRKUNGEN KUNDENINDIVIDUELLER MARKETINGKONZEPTE (1997).

[29] For the details on this argument of supply side substitutability, see Commission Notice 1997 O.J. (C 372/5), PP 20-23.

[30] *See* Michael Anderson & Susan Leigh Anderson, *The Status of* ***Machine*** *Ethics: A Report from the AAAI Symposium*, MINDS & ***MACHINES*** 1, 3-4 (2007).

Technological change has traditionally been accompanied by fundamental societal changes, often including massive job losses. [31]For instance, after the completion of the first U.S. transcontinental telegraph line in 1861, the services rendered by Pony Express riders became obsolete. [32]Telegraph lines, however, soon became the basic fundament for the emergence of the new telecommunication industry, creating a myriad of new jobs over time. The increasing ***use*** of AI indeed poses the question of whether AI can be seen as the new **[*186]** telegraph line, creating 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. [33]There are significant uncertainties over whether a more automated, digital society and economy will leave sufficient opportunities for people to earn a livelihood. [34]While precise calculations are still lacking, some studies conducted estimate that 49% of activities ***used*** in jobs in the global economy [35]and between 22% and 44% [36]of jobs in the developed world could be lost as a consequence of an increasingly digitalized and automated economy. The STOA study conducted by the European Parliament Research Service in 2016, however, appears to be more optimistic and presents a more differentiated outlook. While this study predicts a loss in the number of jobs in the fields of agriculture, food supply [37]and transportation, [38]it predicts that other sectors will likely only see a change in the type of jobs, including a rise in the number of highly skilled jobs in relation to services rendered (e.g., for disabled and elderly people). [39]Generally speaking, the more a job requires social **[*187]** intelligence, the less likely it is that such job will be computerized. [40]A recent study conducted in the UK estimates that countervailing displacement and income effects are likely to broadly balance each other out over the next twenty years. [41]

2. Liability for Damages Caused by AI Systems

---

[31] For a description of the challenges associated with the increasing ***use*** of computers see Keith Abney et al., *Robot Ethics: Mapping the Issues for a Mechanized World*, 175 ARTIFICIAL INTELLIGENCE 942 (2011).

[32] MICHAEL J. QUINN, ETHICS FOR THE INFORMATION AGE 24 fig.1.12 (7th ed. 2017).

[33] *See* European Parliament Resolution of 16 Feb. 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), *http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN* [hereinafter European Parliament].

[34] GERMAN FED. MINISTRY OF EDUC. & RESEARCH (BUNDESMINISTERIUM FÜR BILDUNG UND FORSCHUNG), ZUKUNFTSMONITOR IV: WISSEN SCHAFFEN - DENKEN UND ARBEITEN IN DER WELT VON MORGEN 3-6 (2017). 58% of a group of 1,004 participating German citizens believed that digitalization and robotics will cause more job losses than create new jobs. *Id*. at 3. 80% believed that the main part of routine jobs will be done by ***machines*** or computer programs in the year 2030. *Id*. at 4. 81% expect that due to the speed of technological change more and more people will become increasingly isolated. *Id*. at 6.

[35] MCKINSEY & CO., A FUTURE THAT WORKS: AUTOMATION, EMPLOYMENT, AND PRODUCTIVITY 5 (2017).

[36] RICHARD BERRIMAN & JOHN HAWKSWORTH, PRICE WATERHOUSE COOPERS, WILL ROBOTS STEAL OUR JOBS? 1 (2017). The potential impact of automation on the UK and other major economies suggests that up to 30% of UK jobs could potentially be at high risk of automation by the early 2030s, while figures differ for other economies (US: 38%, Germany: 37%, Japan: 24%). *Id*. at 16.

[37] EPRS, *supra* note 5, at 23.

As AI systems are ***used*** more frequently in close proximity to humans, it is important to determine who should be held liable for eventual damages caused by the operation of AI systems. [42]This is even more relevant as a malfunction in automated systems may have multiplying effects.

The critical ***ethical*** issue is whether a human being should be ***responsible*** for damages caused by an AI-driven or otherwise automated ***machine***, which after consideration of data has taken an autonomous decision that caused harm to human life, health or property. While one could argue that the person--having implemented or made ***use*** of the AI system in fulfillment of an owner obligation--is ***responsible***, this question will become more critical as the decisions taken by AI systems become more autonomous. Legal accountability is generally not a given if independent events or decisions cause a specific damage, unless the law provides for strict liability regimes as is the case in European product liability law. [43]Merely fault-based liability regimes might, therefore, expose victims of AI-caused damages to significant protection gaps.

[*188] It is debatable whether the existing mixture of fault-based damages compensation regimes and strict liability rules on product liability are appropriate for the potential harm caused by AI and autonomous systems. [44]The concepts of responsibility, accountability and liability, consequently, are some of the fundamental legal and ***ethical*** concerns that need to be discussed in depth in relation to new AI applications. [45]It is of utmost importance to critically review the concept of autonomy. As the technology stands today, even AI-driven ***machines*** are still programmed by humans and work within the limits of the respective human-made programming. Accordingly, it does not seem to be the right approach to consider an AI-driven decision as a truly autonomous decision which would protect from liability the person who programmed, ***used*** or manufactured the AI. [46]This may, however, change when AI technology advances.

3. Lack of Transparency of AI

---

[38] Note this designates discussing the replacement of standard taxis by driverless cabs a security and safety issue, but it is also relevant to the transport sector. *Id*. at 53.

[39] *Id*. at 10. For a more differentiated and balanced approach to the evaluation of potential impacts of AI on employment and jobs, see IEEE GLOB. INITIATIVE ON ETHICS OF AUTONOMOUS & INTELLIGENT SYS., ETHICALLY ALIGNED DESIGN - A VISION FOR PRIORITIZING HUMAN WELL-BEING WITH AUTONOMOUS AND INTELLIGENT SYSTEMS, 136 (2017), http://standards.ieee.org/develop/indconn/ec/ead_v2.pdf [hereinafter IEEE]. For an analysis of the susceptibility to computerization of different types of jobs, see CARL BENEDIKT FREY & MICHAEL A. OSBORNE, THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERIZATION (2013), https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.

[40] FREY & OSBORNE, *supra* note 39, at 27, 40.

[41] PRICEWATERHOUSECOOPERS, UK ECONOMIC OUTLOOK 49 (2018), https://www.pwc.co.uk/economic-services/ukeo/ukeo-july18-full-report.pdf.

[42] Communication Commission on Liability for Emerging Digital Technologies SWD (2018) 137 final [hereinafter Liability for Emerging Digital Technologies]; Artificial Intelligence for Europe, *supra* note 26, at 15, 16; EPRS *supra* note 5, at 8.

[43] For European law, see in particular 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.

[44] According to a European Commission consultation of 2017, 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*

Another growing criticism is the lack of transparency within AI systems. [47]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***." [48]Yet, **[*189]** transparency is necessary for different reasons. [49]From a user perspective, transparency is important in order 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. [50]If an AI system does something unexpected, the user, at least, needs to be able to be informed of the reasons and parameters considered by the AI system.

Further, transparency is necessary in case of harm caused by AI systems so that an investigation of the respective accident may take place. In order to allocate responsibility to the relevant person or entity, Courts, lawyers and expert witnesses need to be in an appropriate position to understand why and how an AI system has taken certain decisions and actions. Finally, if the ***use*** of certain AI agents should be subject to marketing authorization or other approval procedures, competent authorities need to understand the functioning of such algorithmic agents. Otherwise, they would not be able to evaluate the risks associated with the operation of the relevant system. This need is already evident to the extent that AI systems are ***used*** for pharmaceutical purposes or within medical devices. [51]For example, the FDA has already issued the first approval for a smart drug with an ingestible sensor embedded in a pill, which records that the medication was taken by the patient. [52]

[*190] *4. Loss of Humanity in Social Relationships and Lack of Protection of Human Life and Human Dignity*

Even more critical than possible job losses and liability issues, 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.
>
> . . . .

---

*(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 a first analysis, see European Commission, SWD (2018) 137 final, in particular nos. 2 and 4.

[45] IEEE, *supra* note 39, at 148; European Parliament, *supra* note 33 at rec. 49 et seqq.

[46] Gerald Spindler, *Zivilrechtliche Fragen beim Einsatz von Robotern, in* ROBOTIK IM KONTEXT VON RECHT UND MORAL 66 (Eric Hilgendorf ed., 2013).

[47] *See, e.g.*, NICK BOSTROM & ELIEZER YUDKOWSKY, THE ETHICS OF ARTIFICIAL INTELLIGENCE, 1 (2011), *available at* https://intelligence.org/files/EthicsofAI.pdf. The lack of transparency is in particular due to the technical design of deep ***learning*** mechanisms, see*infra* section I.1.a.

[48] INFO. COMM'R'S OFFICE, *supra* note 16, at 10. For issues related to the black box effect in AI algorithms ***used*** for medicinal purposes, see W. Nicholson Price II, *Black Box Medicine*, 28 HARV. J.L. & TECH. 419, 432 (2015).

[49] To understand the purpose for which IEEE P7001 TM standard was developed see *P7001 - Transparency of Autonomous Systems*, IEEE STANDARDS ASS'N, https://standards.ieee.org/develop/project/7001.html (follow "Approved Pars" hyperlink) (last visited 6 Oct. 2018).

[50] Bostrom & Yudkowsky, *supra* note 47, at 1.

[51] For the regulatory approval mechanisms applicable to pharmaceuticals, see Commission Regulation 726/2004 of 31 March 2004, Laying Down Community Procedures for the Authorisation and Supervision of Medicinal Products for Human and Veterinary ***Use*** and Establishing a European Medicines Agency, 2004 O.J. (L 136) 1; Council Directive 2001/83, 2001 O.J. (L 311) 67 (EC); Council Directive 2001/82, 2001 O.J. (L 311) 1 (EC); in relation to medical devices see Commission Regulation 2017/745 of 5 April 2017 On Medical Devices, 2017 O.J. (L 117) 1.

> We've already started putting aside our feelings, intuitions and dreams in favour of more reasonable choices, calculated by an algorithm and powered by objective data . . . . [53]

In addition, more automation and reliance on AI for making decisions in our daily lives may lead to a decrease in social contacts. Indeed, increased man-to-***machine*** interaction may result from AI applications such as healthcare robots in hospitals, service robots for elderly people, service robots ***used*** in the field of tourism and--last but not least--AI enabled toys. It is entirely unclear how these developments might affect our emotional life and ways of thinking. [54]Even typical human strengths such as emotions and intuition could be affected significantly by the increasing reliance on AI for decision-making purposes. [55]The new technological developments around the implementation and ***use*** of AI will consequently give rise to fundamental questions such as 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***. A further issue arising in relation to AI systems that are ***used*** for social interaction is how such systems should behave from an ***ethical*** and moral point of view and to what extent self-***learning*** mechanisms and autonomous behavior should be allowed. [56]

[*191] *5. Loss of Privacy*

An additional concern is the loss of privacy associated with AI. In order to make intelligent decisions, AI systems need to collect and process data. Thus, access to data is of fundamental importance for the further development of digital technologies in general, and AI in particular. [57]In certain societies, protection and maintenance of privacy in data is a major ***ethical*** concern. [58]In such societies, it is considered crucial to make sure that while accessibility of non-personal data is improved, there are sufficient data protection standards to protect personal data. [59]From a European perspective, the General Data Protection Regulation, a new and stricter regulatory framework regarding the ***use*** of personal data, became effective on May 25, 2018. [60]

---

[52] *See* Press Release, Food & Drug Admin., FDA Approves Pill with Sensor that Digitally Tracks if Patients Have Ingested Their Medication (Nov. 13, 2017), https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm584933.htm.

[53] Bernard Cathelat, *How Much Should We Let AI Decide For Us?, in* HUMAN DECISIONS: THOUGHTS ON AI 132, 134 (2018), http://unesdoc.unesco.org/images/0026/002615/261563e.pdf.

[54] Abney, *supra* note 31, at 942.

[55] Olaf Groth et al., *Rules for Robots: Why We Need a Digital Magna Carta for the Age of Intelligent* ***Machines****, in* INTERNATIONAL REPORTS 16, 18 (2018), http://www.kas.de/wf/doc/kas_52115-544-2-30.pdf?180418140416.

[56] EPRS, *supra* note 5, at 8.

[57] JOSEPH DREXL 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" 3 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2959924; Artificial Intelligence for Europe,*supra* note 26, at 10.

[58] *See* European Parliament, *supra* note 33 (emphasizing the European Union legal framework must be complied with in the areas of robotics in order to respect the right to the protection of personal data); EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 61 (2014); *Algorithms: How Companies' Decisions About Data and Content Impact Consumers: Hearing Before the Subcomm. On Dig. Commerce & Consumer Prot. of the H. Comm*. *On Energy & Commerce*, 115th Congress 24 (2017) (statement of Frank Pasquale, Professor, Univ. of Md.).

[59] Artificial Intelligence for Europe, *supra* note 26, at 10.

[60] Commission Regulation (EU) 2016/679 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, 2016 O.J., (L 119) 1.

Appropriate means and mechanisms must be implemented to protect AI systems against abuse. For instance with connected mobility, manipulation of automobile infotainment systems may eventually even cause traffic accidents. More concretely, an automobile's connected mobility system that is not sufficiently protected against abuse may allow hackers to take remote control of the vehicles while they are in operation. The legal question of what liability a car manufacturer should have if its infotainment system is hacked is a matter of ongoing debate. [61]

[*192] *6. Loss of Personal Autonomy*

The development of intelligent assistants may be convenient and may help to manage administrative and other tasks of daily life. At the same time, 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 potential consequence of digital technology overuse describing the deterioration or breakdown of cognitive abilities. [62]Overuse of digital technology may further impact personal autonomy, depending on the degree of digital assistance increasingly relied upon for the completion of even trivial tasks, like watering indoor plants. [63]As a consequence of the growing reliance on digital assistance, basic human capabilities could get lost. [64]

7. Restriction of Competition and Plurality of Opinions: 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. [65]In view of the rapid development of digital products and markets, such bias multiply quickly and consequently have a widespread impact. [66]The increasing ***use*** of algorithms can even reduce the plurality of views expressed in public discussions. For example, consider the ***use*** of chat bots. Chat bots pick up certain views and facts to share with as many readers as possible. Such automated mass distribution may cause a critical information bias and distort the actually predominant public opinion. This is a particular concern to society if wrong or biased facts (often referred to as so-called "fake news") are intentionally spread by chat bots to influence certain decision-making processes. [67]Corresponding new communication strategies may **[*193]** consequently even be liable for charges of tortious interferences on elections and other democratic decision-making procedures. [68]

---

[61] *Cahen v. Toyota Motor Corp., 147 F. Supp. 3d 955, 967-68 (N.D. Cal. 2015)*; Flynn v. FCA US, LLC, No. 15-cv-0855-MJR-DGW 2016 WL 5341749 at *2 (S.D. Ill. 2016).

[62] MANFRED SPITZER, DIGITALE DEMENZ (2012); Larry Dossey, *FOMO, Digital Dementia, and Our Dangerous Experiment*, EXPLORE, Mar./Apr. 2014, at 69, 70-71; Markus Appel & Costanze Schreiner, *Digitale Demenz? Mythen und wissenschaftliche Befundlage zur Auswirkung von Internetnutzung*, 65 Psychologische Rundschau 1, 8-10 (2014).

[63] *See, e.g.*, Koubachi, ITUNES, https://itunes.apple.com/de/app/koubachi-persönlicher-pflanzenpflege-assistent/id391581160?mt=8 (last visited Oct. 6, 2018).

[64] MANFRED DANIEL & DIETER STRIEBEL, KÜNSTLICHE INTELLIGENZ, EXPERTENSYSTEME: ANWENDUNGSFELDER, NEUE DIENSTE, SOZIALE FOLGEN 103 (1993).

[65] EPSC, *supra* note 6, at 7.

[66] *Id*.

[67] Bernd Holznagel, " *Phänomen, Fake News*" *-Was ist zu tun?*, MMR 18, 19 (2018); Boris Paal & Moritz Hennemann, *Meinungsvielfalt im Internet*, ZRP 76, 77 (2017).

[68] *See* MEG LETA AMBROSE, THE LAW AND THE LOOP (2014) (discussing Congressional concern regarding the rise of robocalls in the late 1980s). *See also* Chuck Todd & Carrie Dann, *How Big Data Broke American Politics*, NBCNEWS (Mar. 14, 2017), *https://www.nbcnews.com/politics/elections/how-big-data-broke-american-politics-n732901*; Maurice Stucke,*supra* note 65, at 1271-79.

In addition to a possible reduction of the plurality of views and opinions, algorithms may also reduce competition and thereby negatively impact innovation. [69]The Department of Justice, for instance, found a group of Amazon marketplace sellers guilty of an antitrust infringement by having designed and shared among themselves dynamic pricing algorithms programmed to act in conformity with their agreement. [70]Corresponding concerns may arise if companies engage in the ***use*** of the same pricing algorithms. ***Using*** the same algorithms could also result in price fixing above the competitive level. [71]

8. 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 also needs to be born in mind that AI-generated decisions and results are based on algorithms ***using*** statistical models by analyzing certain amounts of data. [72]The ***use*** of statistical models, however, may generate wrong decisions and results, either because the data analyzed for a specific case does not accurately reflect the individual circumstances of the respective scenario, because the data analyzed is biased or incorrect, or because the statistical model is incomplete or incorrect. [73]From a legal perspective, decision-making processes relying on statistical models involve an automatic discrimination with regard to these cases that differ from the statistical role model. [74]

[*194] Further, computer and software technology is susceptible to errors and manipulation. [75]Even computer and software systems believed to be secure, like the network of the government of the Federal Republic of Germany, have already been hacked successfully. [76]The German Federal Office for Information Security (BSI) concluded in its report on the State of IT Security in Germany 2017 that "the risk situation is continuously tense and at a high level." [77]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

---

[69] *See* Org. for Econ. Co-operation & Dev., *Algorithms and Collusion*, (June 21-23, 2017), https://one.oecd.org/document/DAF/COMP(2017)4/en/pdf (discussing the risk algorithms pose to competition and effects of policy choice with respect to regulating algorithms on innovation).

[70] *Id*. at 27.

[71] *Id*.

[72] *See infra* Section I.1.a.

[73] HOFSTETTER, *supra* note 3, at 361 ("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.").

[74] The German Federal Supreme Court stated that extrapolating from statistical data to individual cases poses general difficulties and that it is generally impossible to make a decision, based on statistical data, whether a result of a specific assessment is correct. Bundesgerichtshof [BGH] [Federal Court of Justice], Dec. 17, 1998, NEUE JURISTISCHE WOCHENSCHRIFT [NJW] 657, 658-61, (Ger.).

[75] *See* Bostrom & Yudkowsky, *supra* note 47, at 2; ONE HUNDRED YEAR STUDY ON ARTIFICIAL INTELLIGENCE, ARTIFICIAL INTELLIGENCE AND LIFE IN 2030 42, (2016), https://ai100.stanford.edu/sites/default/files/ai100report10032016fnl_singles.pdf. For examples of computer and software systems that are susceptible to errors or manipulation, see FED. OFFICE FOR INFO. SEC., THE STATE OF IT SECURITY IN GERMANY 2017 14-16 (2017), https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Securitysituation/IT-Security-Situation-in-Germany-2017.pdf?__blob=publicationFile&v=3 (regarding the possible manipulation of traffic lights or of "smart home components" that regulate for access control in advance of a burglary).

over systems." [78]This indicates that software and hardware systems that are also the basis of AI are highly error-prone and susceptible to manipulation.

9. Manipulation, Surveillance and Illegal Behavior

Finally, AI involves a high risk of being abused for manipulation, surveillance, or other quasi-legal purposes. For instance, democratic elections may be manipulated, [79]and facial recognition systems may be abused to control citizens. [80]Companies **[*195]** may further ***use*** price determination algorithms to agree on sales prices above market level and thereby harm consumers. [81]

*C. Specific Benefits and Risks Related to the **Use** of AI for Healthcare Purposes and Assisting Elderly People*

This section discusses the specific benefits and risks associated with AI in healthcare, given the prevalence of AI in this industry. In relation to healthcare, AI systems such as surgery robots and telemedicine (i.e., medical devices that can assist patients at home) provide obvious advantages. Surgery robots may be more accurate and less susceptible to personal and environmental performance issues. Telemedicine allows patients to be monitored at home by collecting real-time data of their health conditions, potentially significantly reducing hospital stays. [82]Reduced hospital stays would reduce the patients' risk of catching further infections. By allowing patients to recover at home, telemedicine may also reduce the time for their convalescence. In addition, the availability of medical assistance in rural areas and developing countries may be improved. [83]

AI systems may also be suited to considering the individual particularities of patients and thereby fostering individualized patient treatment methods. An example is the increasing ***use*** of 3D printing technologies that can be ***used*** to fabricate tailor-made body part prosthetics. [84]This may again benefit patients' health and reduce the time for convalescence.

Ultimately, an AI-driven healthcare system ***using*** digital technology and smart home caring devices could even lead to a shift in the focus of the current healthcare systems towards preventive care. [85]All of these trends--a

---

[76] *Hacker Drangen in Deutsches Regierungsnetz Ein*, ZEIT ONLINE (Feb. 28, 2018), http://www.zeit.de/digital/datenschutz/2018-02/hacker-dringen-in-deutsches-regierungsnetz-ein.

[77] FED. OFFICE FOR INFO. SEC., *supra* note 75, at 75.

[78] *Id*.

[79] Vyacheslav Polonski, *Artificial Intelligence Can Save Democracy Unless It Destroys It First*, OXFORD INTERNET INST. (Aug. 10, 2017), https://www.oii.ox.ac.uk/blog/artificial-intelligence-can-save-democracy-unless-it-destroys-it-first/.

[80] Brad Smith, *Facial Recognition Technology: The Need for Public Regulation and Corporate Responsibility*, MICROSOFT (July 13, 2018), https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/.

[81] ORG. FOR ECON. CO-OPERATION & DEV., ALGORITHMS AND COLLUSION: COMPETITION POLICY IN THE DIGITAL AGE, 18-21 (2017), http://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm.

[82] For a more detailed description of telemedicine ***use*** cases, see Deborah Lupton, *Digital Health Technologies and Digital Data: New Ways of Monitoring, Measuring and Commodifying Human Embodiment, Health and Illness, in* RESEARCH HANDBOOK ON DIGITAL TRANSFORMATIONS 85 (2016), http://ssrn.com/abstract=2552998.

[83] *Id*. at 85.

[84] *Id*. at 87.

[85] *See* MCKINSEY GLOB. INST., ARTIFICIAL INTELLIGENCE: THE NEXT DIGITAL FRONTIER? 61 (2017).

more prevention-based medical system, reduced hospital stays, precision medicine, and reduced convalescence--may not only improve people's health but also significantly reduce public healthcare costs.

Applications in this field may become all the more relevant in view of an aging society and increasing life expectancy, which will **[*196]** result in a fundamentally different balance between generations within society. [86]Elderly people may further benefit from smart home applications, wearable sensors, and robots, as these devices could assist them in their daily lives. [87]For instance, the elderly could ***use*** AI to monitor their health conditions and call for medical help as soon as the need arises. To go one step further, AI could even take over the decision-making power of a distressed individual and could call for medical help irrespective of a user's individual consideration and will.

These possible advantages to AI within the healthcare realm are counteracted by ***ethical*** concerns. In addition to the general concerns resulting from less personal interaction between humans, such scenarios give rise to the following fundamental ***ethical*** issues: Who decides the decision-making power of a particular AI system, and what should the level of autonomy of such a system be? Should AI decide in a paternalistic manner so that it can override the user's will if this were deemed to be detrimental for the user's health? Who is liable if an AI-driven decision is wrong and damages a user's health? If an AI system monitoring its user permanently collects an extensive amount of data that could be of interest for burglars and other criminals, who is ***responsible*** for making sure that a respective AI system is not susceptible to being hacked? How can it be guaranteed that a user's data is only accessible to persons authorized by the user? How can it be guaranteed that the considerations taken into account by an AI system can be traced back for the purpose of allocation of liability?

Several approaches exist for how to address these ***ethical*** concerns regarding the implementation of AI in healthcare. One idea is that AI systems should be designed so that they can always show their human user the registered process which led to their actions; this would permit the identification of any sources of uncertainty and show any assumptions the AI relied upon. [88]Another proposal is to invite AI system designers to consider adopting an identity tag standard. [89]Under such a standard, no AI system would be released without an identity tag in order to maintain a clear line of legal accountability. [90]Moreover, the industry could consider implementing a standard that requires any and all AI systems be equipped with a specific technology that allows for an immediate stop of all operations of the system. [91]Ultimately, the industry could agree on a **[*197]** certain level of autonomy to be implemented in AI systems for elderly people and provide for a technology which makes sure that a user's will can at all times override an AI-driven decision. AI could accordingly be programmed in a standardized manner guaranteeing that it always has to take a user's latest will into account. [92]

In light of these approaches, it is ultimately up to legislatures to decide on the allocation of liability and responsibility. Basic models that could be applied are either to hold the user of AI liable to the extent to which he

---

86 European Parliament, *supra* note 33, Introduction, paragraph F.

87 *See* European Parliament, *supra* note 33, paragraphs 31.

88 IEEE, *supra* note 39, at 159.

89 *Id*. at 155.

90 *Id*.

91 Often referred to as the so called "kill switch." *See, e.g., Google Developing Kill Switch AI*, BBC NEWS (June 8, 2016), http://www.bbc.com/news/technology-36472140. We would advocate against ***using*** this terminology, however, because it creates the impression that AI is something close to human life-something it is not, and something it should never be considered similar to.

92 This is similar to the requirements which need to be complied with in order for a patient decree (so called "Patientenverfügung" in German) to be binding upon a medical doctor. For the criteria under German law, see BÜRGERLICHES GESETZBUCH [BGB] [CIVIL CODE] § 1901a para. 1.

makes ***use*** of the AI in order to complete his own task [93]or to establish a regime of strict liability to be borne by the manufacturer, owner, or operator of the AI system in question. [94]Alternatively, one could consider new laws introducing the concept of an "e-person" on the basis of which an AI system would be held directly liable. [95]This would, however, require the establishment of an appropriate financing or insurance system to make sure that AI systems are sufficiently capitalized and cannot be abused as a potential way to circumvent liability. [96]

As an interim result, even this very brief look at the potential ***use*** of AI for healthcare purposes makes it clear that the many risks and concerns that may arise cannot be resolved by one uniform approach. Instead, this example underlines the lack of a general answer to the question about which mechanism should be ***used*** for the implementation of ethics into AI systems. While certain ***ethical*** considerations can only be dealt with on a regulatory basis (e.g., the question of how liability and responsibility for damages caused by AI will be dealt with, or more basic questions like whether and to which extent AI constitutes a permissible technology to be ***used*** in a certain respect), others are more amenable to an implementation by setting **[*198]** industry standards (e.g., the requirement of a control switch to make sure that AI system can be switched off at any time and the requirement that an AI system must keep a log of all of its actions and considerations). A successful implementation of ethics into AI systems, therefore, requires a mix of mechanisms and accordingly an in-depth coordination and discussion between the various stakeholders. The possible solutions will be discussed in more detail in the following sections.

III. MEANS TO IMPLEMENT ETHICS IN AI APPLICATIONS

The potential benefits of AI create a need to mitigate or, when possible, even rule out risks and other ***ethical*** concerns, so we can best ***use*** the technology. This Article intends to contribute some ideas on the development of an AI governance regime and how ***ethical*** decision-making processes can be implemented in specific AI systems from a policy-making perspective. To this end, and in order to take into account the multitude of potential ***ethical*** conflicts that may arise in the course of AI operations, the following section intends to review a variety of potential regulatory approaches. Technical solutions, as well as traditional regulatory approaches, will be considered including binding and non-binding measures of self-regulation. While technical solutions are directly implemented into an AI-driven product, regulatory approaches oblige manufacturers and/or users of such products to ensure that certain normative standards are complied with. In each respect, the corresponding benefits and drawbacks will be highlighted.

A. Technical Means and Mechanisms: Ethics Compliance by Design

Irrespective of potential regulatory approaches, it is necessary to think about how to construct AI systems technically in a way that such systems *per se* behave in an ***ethical*** manner, at least in specific critical situations ("ethics compliance by design"). This section will look at how AI systems can be programmed to behave ethically. This section starts with an overview of possible technical approaches to implement ***ethical*** decision-making principles into AI through bottom-up and top-down approaches. It then explains casuistic as well as dogmatic approaches. This will be followed by the concept of implementing a guardian AI. The section concludes that technical means, even though possible to a certain extent, are not sufficient to provide for the maintenance of ***ethical*** decision-making processes in a more automated and AI driven world.

[*199] *1. Overview on Technical Approaches to Implementing Ethics into AI*

---

[93] Such a concept would be similar to the German concept of liability for acts committed by vicarious agents (so called "Erfüllungsgehilfen" in German) pursuant to BÜRGERLICHES GESETZBUCH [BGB] [CIVIL CODE] § 278.

[94] Such a concept would be similar to the European concept of product liability as established by the Product Liability Directive, Council Directive 85/374/EEC of 25.07.1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, 1985 O.J. (L 210/29).

[95] Christiane Wendehorst, *Die Digitalisierung und das BGB*, NJW 2609 (2016); Bräutigam & Klindt, NJW 1137, 1138. (2015).

[96] *See also* IEEE, *supra* note 39, at 148.

i. Bottom-up Versus Top-down Approaches: The Tay Example

To implement ***ethical*** decision-making criteria technically, bottom-up or top-down approaches are possible. [97]***Using*** a bottom-up approach, ***machines*** would be expected to observe human behavior 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 only what is common. [98]In 2016, only shortly after its launch, Microsoft's chat bot Tay started making racist, inflammatory, and political statements which had been taught to it by users determined to undermine it. [99]

Therefore, it appears that from a technical perspective a top-down approach is better suited to implement ethics into AI. Under a top-down approach, ***ethical*** principles would be programmed directly into an AI system. [100]In the field of predictive policing, for instance, a sentencing algorithm could be programmed to ensure compliance with legal and ***ethical*** non-discrimination requirements. It could operate in a manner ensuring that racial-specific data is in no way ***used*** for making a social prognosis on the basis of which judges decide on whether or not a specific sentence can be suspended or not. A stricter top-down approach could even prohibit the ***use*** of AI for making prognostic judicial decisions.

ii. Casuistic Versus Dogmatic Approaches

Ethical principles could also be implemented in AI systems on a casuistic or dogmatic basis. Under a casuistic approach, ***machines*** would be programmed as to how to react specifically in each situation where they may have to take an ***ethical*** decision. For example, consider a healthcare robot that could be programmed to always consider the will of its user (i.e., the patient) before taking a specific action. If no clear will was previously expressed by the user 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 **[*200]** 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 state of health of the user. Difficulties would, however, arise when no advance directive is available, and the user is not in a position to express its will anymore. Probably, in consideration of human life being protected as an absolute fundamental right, [101]a default setting of the AI system in such a scenario should be take action that has the highest probability of saving the user's life.

Second, rather than anticipating all possible scenarios where an AI system would need to take an ***ethical*** decision and programming the AI system (like in the casuistic approach), AI could be programmed under a dogmatic approach. Under a dogmatic approach, systems could be programmed in line with a specific ***ethical*** school of thought--such as utilitarianism, Kantian ethic, [102]Asimov's Three Laws of Robots, [103]or the Golden

---

97 Amitai Etzioni & Oren Etzioni, *Incorporating Ethics into Artificial Intelligence*, 17 J. ETHICS 2017, sec. 1.2.; COLIN ALLEN ET AL., ARTIFICIAL MORALITY: TOP-DOWN, BOTTOM-UP, AND HYBRID APPROACHES 150 (2005).

98 Etzioni, *supra* note 97.

99 Elle Hunt, Tay, *Microsoft's AI chatbot gets a crash course in racism from Twitter*, THE GUARDIAN (Mar. 24, 2016, 2:41 PM), https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

100 Etzioni, *supra* note 97.

101 G.A. Res. 217 (III) A, Universal Declaration of Human Rights, art. 3, (Dec. 10, 1948); Charter of Fundamental Rights of the European Union, 2000 O.J. (L, 364/1).

102 Jessica Heesen, *Mensch und Technik. Ethische Aspekte einer Handlungspartnerschaft zwischen Personen und Robotern, in* ROBOTIK IM KONTEXT VON RECHT UND MORAL 281 (Eric Hilgendorf ed., 2013).

103 *See also* JOHN FRANK WEAVER, ROBOTS ARE PEOPLE TOO 4 (2014) ("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

Rule--[104]which requires that one should not treat others in a way that one would not like to be treated oneself and which as such is the basis of many international philosophies and different religions.[105]A major issue with this idea, however, is that such an approach would blindly follow that specific school--making it a quite drastic approach. Further, blindly following a specific school may result in a decision that is, in a specific scenario, unethical. Most ethicists apply rules of different schools of thought to resolve a specific ***ethical*** issue in order to take well-balanced decisions, rather than just applying a single **[*201]** doctrine of thought.[106]Moreover, it is not yet clear whether AI systems could be so programmed to singularly follow a specific school.

Therefore, it appears--at least for the time being--that the preferable technical approach for programming ***ethical*** principles into AI systems is to do so on a more casuistic basis, relying on specifically programmed decision-making structures. Still, it remains the AI system designers' challenge to generally deal with this question and decide on which design philosophy they choose for algorithmic decision-making frameworks. As a potential approach to resolve the issue of situation-specific ethics application, it is suggested that ***ethical*** requirements for computational systems should be developed collaboratively and in a sufficiently transparent manner. To this end, an ***ethical*** protocol on the basis of which the designer's explicit ***ethical*** principles can be reviewed should be established. Such ***ethical*** protocols can then be referred to in order to achieve consistency in the decision-making process.[107]For this purpose, close cooperation between researchers, developers and policy-makers is necessary in order to develop a common understanding of the relevant ***ethical*** principles on the basis of which the "good AI society" shall be developed.[108]

2. Implementing AI on a Technical Meta-level

In view of the autonomous nature of decisions taken 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 unlawful or unethical decision taken by the basic AI to an appropriate enforcement authority or agency.[109]These requirements and benefits can be transformed into concrete technical solutions when they are available.[110]

---

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.").

[104] *See, e.g.*, 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 BRIAN LEPARD, HOPE FOR A GLOBAL ETHIC 35 (2005); DIE "GOLDENE REGEL" IN DEN WELTRELIGIONEN, https://www.erzdioezesewien.at/dl/OKrlJKJlMnklJqx4kJK/11JKW_Goldene_Regel_Zivlilcourage_konkret_-_Schulmodul.pdf (last visited Oct. 6, 2018).

[105] Colin Allen et al., *Why* ***Machine*** *Ethics?*, 21 IEEE INTELLIGENT SYSTEMS 12, 14. (2006). There is no consensus for how to "practically relocate the social and ***ethical*** duties displaced by automation." MITTELSTADT ET AL., *supra* note 5, at 12.

[106] HEESEN, *supra* note 102, at 282.

[107] MITTELSTADT ET AL., *supra* note 5, at 12.

[108] MITTELSTADT ET AL., *supra* note 5, at 13; CATH ET AL., *supra* note 8, at 20.

[109] Regarding the establishment of a corresponding agency, see Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 395 (2016).

[110] Etzioni & 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." Etzioni, *supra* note 97, at abstract. However, this is not convincing as the compliance of AI systems with ***ethical*** requirements requires a technical implementation of the corresponding requirements, all the more when ***machines*** act increasingly without direct human control.

[*202] However, one of the difficulties of this approach can be demonstrated by the following scenario: With regard to autonomous driving, each time an autonomous car is driven above the speed limit, its AI Guardian reports the infraction. If the standard AI is merely copying the driving of human drivers in that setting, how is the correct punishment to be assessed? Indeed, in some cases it would be more dangerous not to speed, and thus, how should the AI deal with them? Likewise, the risk of dysfunctional decisions is significant when it comes to questions around social morality. For instance, in the field of criminal prosecution, a Guardian AI may face difficulties when it comes to sentencing and deciding whether or not a sentence shall be suspended on probation or not. A major question in this regard is whether and which criteria such as criminal records, type of offence, amount of damage caused, stability of social relationships, responsibility for the maintenance of children, parents or other family members shall be relied upon. When and in which cases should such criteria be referred to and who takes the responsibility for that decision? A Guardian AI following a specific pre-determined technical protocol might not be sufficiently flexible and empathetic to make the right decision.

3. The Insufficiency of Technical Means and Mechanisms

While a technical approach, whether casuistic or dogmatic, may nevertheless eventually be a possible means to resolve ***ethical*** issues in certain cases, such an approach is probably in any event insufficient to ensure that AI systems do indeed take into account ***ethical*** considerations for their decision-making process. AI systems are constructed by, need to be programmed by and will be ***used*** by humans and companies. Therefore, unless the persons and companies ***responsible*** for programming and ***using*** an AI system 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 beneficial for them, economically or otherwise. To make sure that AI systems behave according to ***ethical*** principles, it is therefore necessary to adopt a variety of regulatory mechanisms, including binding legal requirements or creation of economic incentives, to promote ethically aligned AI system design.

[*203] *B. Regulatory Approaches: Policy-Making Instruments*

Considering the insufficiency of technical means for the purpose of ensuring ***ethical*** AI decision-making processes, it is necessary to look to traditional regulatory approaches, via policymaking. The potential approaches--each discussed in turn--include legislation, international resolutions and treaties, bilateral investment treaties, self-regulation and standardization, certification, contractual rules, soft law, and agile governance. These can be referred to for the purpose of ensuring ***ethical*** compliance by those persons and companies constructing, selling, and ***using*** AI-driven ***machines*** and autonomous systems and thereby establish a human centric AI governance regime.

1. Legislation

i. Pros and Cons of Legislation

Legislation is the typical regulatory approach to implement ***ethical*** rules such as the primacy of the user's will, the obligation not to harm other persons, and the obligation not to destroy other people's belongings. The advantage of legislation is that it provides for binding and enforceable rules that are established and consequently generally accepted on the basis of a democratic process ensuring transparency and participation of the people and relevant interest groups. Additional advantages are that the process of establishing legislation is subject to the rule of law, and legislation established within a democratic process is transparent. In certain contexts, legislation provides for at least a certain level of legal certainty and social acceptance.

However, legislation, even in democracies, also has some shortcomings. Due to the democratic lawmaking process and in light of corresponding compromises having to be made, laws often only protect a minimum consensus of ***ethical*** rules. Legislation may therefore not be an appropriate regulatory instrument insofar as specific ***ethical*** interests of selected individuals are concerned. At the same time, however, it has to be born in mind that laws may typically become necessary in order to protect interests and concerns of specific minorities. An additional significant disadvantage of legislation is the territorial limitation; laws, basically only bind people of and

within the respective national states. [111]Also, the democratic lawmaking process is usually complex, lacks flexibility and therefore tends to be relatively slow. Therefore, it is often **[*204]** difficult to respond to technical developments and corresponding regulatory needs quickly. Finally, legislation is often perceived as impacting innovation negatively. If true, this may ultimately put domestic businesses at a disadvantage in comparison to businesses residing in less regulated countries.

However, in certain circumstances, legislation may incentivize innovation as companies need to compete to adopt compliant technologies and business models. For instance, with regard to data protection, efficient legislation may even be considered to be 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. [112]The reason is that a strict level of data protection assures a higher level of trustworthiness on the side of consumers. Business models complying with correspondingly stricter standards of data protection are therefore more likely to be accepted by potential users. This consideration should also be born in mind in relation to other ***ethical*** rules and values. Customers might generally welcome 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 potential anti- as well as pro-competitive effects of legislative regulation.

ii. When the Principle of Democracy May Require Legislative Regulation

From a policy-making perspective, whether legislation is chosen as an instrument for regulating ***ethical*** concerns depends on an overall view and balancing of all relevant aspects of the specific ***ethical*** concern, the specific ***use*** case of an AI-driven ***machine*** or automated system and the possible impacts of regulation. In some situations, legislation should be mandatory to protect people from **[*205]** potentially harmful products and technologies. [113]In relation to facial recognition technology, Microsoft's President Brad Smith stated: "While we appreciate that some people today are calling for tech companies to make these decisions -- and we recognize a clear need for our own exercise of responsibility . . . we believe this is an inadequate substitute for decision making by the public and its representatives in a democratic republic." [114]German law generally acknowledges that all questions of a fundamental nature have to be taken by the parliament and should not be surrendered to other policy-makers (so called " *Parlamentsvorbehalt*"). [115]The principle of *Parlamentsvorbehalt* ensures that rules governing such questions of a fundamental nature are established by following a formal procedure characterized by transparency,

---

111 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* 2012 O.J. (L 326/1).

112 EPSC STRATEGIC NOTES, *supra* note 6, at 6. With regard to the data protection standards as established by the EU General Data Protection Regulation, *see supra* note 60. As an interesting 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.

113 For details regarding the regulation of new technologies, see Everhard Holtmann, Parlamentslehre 433 (Raban Graf von Westphalen, vol. 2, 1996).

114 Smith, *supra* note 80.

115 13 Hans D. Jarass & Bodo Pieroth, Grundgesetz für die Bundesrepublik Deutschland art. 20, rec. 47 (2018); HOLTMANN, *supra* note 113, at 439.

that provides for the participation of the parliamentary opposition, and provides the opportunity for the concerned persons and the public to voice their opinion. [116]

iii. Examples for Legislative Regulation in the Field of AI

In many states governed by the rule of law, many existing laws can be applied to AI-driven technology. The remaining decisive question is whether the purposes of existing laws are well-suited to the particularities of AI, such as its lack of transparency, self-***learning*** capabilities, propensity for error, and possible impacts on society and its fundamental values.

a. Application of Existing Laws

(1) Example 1: Data Protection Law

Data protection laws intend to ensure that personal data can only be ***used*** subject to the respective person's prior consent. The underlying ***ethical*** value protected in this regard by the law is a person's right to privacy and private autonomy. These rights, being based on the general right of personality, are fundamental. Consequently, they are considered to require regulation and protection by legislation in view of the aforementioned principle of *Parlamentsvorbehalt*. [117]The German Federal Constitutional Court expressly decided on December 15, 1983 that, in particular with regard to automated data processing, data protection legislation was necessary to protect the so-called right to informational self-determination: "In view of the threats . . . that arise from the ***use*** of **[*206]** automated data processing, the legislature must more than was the case previously, adopt organizational and procedural precautions that work counter to the threat of violation of the right of personality." [118]This ruling applies *a fortiori* in a more and more digitalized environment and, in particular, with regard to deep ***learning*** based AI systems. AI related policy-making must consider this increased need for the protection of the general right of personality to ensure that relevant technologies provide the appropriate means to guarantee the mandatory data protection level. [119]

(2) Example 2: Liability

With regard to liability for damages caused by AI-driven technology, the European Commission has already started to review the existing legal framework and assess to what extent there may be a need for eventual changes to address the challenges posed by AI. [120]Indeed, a careful evaluation is always required to decide whether additional, new, or AI-specific legislation is necessary. The German Federal Supreme Court's Robodoc decision shows how existing rules can often be applied to new technologies. At issue was whether and under which circumstances a medical doctor may be liable for damages caused by the ***use*** of a computer assisted milling process (called "Robodoc") for the implantation of a cementless hip joint prosthesis. According to the court, the ***use*** of a new medical procedure is indispensable. However, to sufficiently protect the private autonomy of patients, new medical techniques may only be applied to a patient if the patient was unequivocally informed that the ***use*** of the

---

116 Bundesverfassungsgericht [BVerfG] [Federal Constitutional Court] Apr. 8, 1997, Neue Juristische Wochenschrift [NJW] 1997, 1975 (Ger.).

117 Gurlit, Neue Juristische Wochenschrift [NJW] 2010, 1035, 1038 (Ger.).

118 Bundesverfassungsgericht [BVerfG] [Federal Constitutional Court] Dec. 15, 1983, Neue Juristische Wochenschrift [NJW] 1984, 419 (Ger.). A free translation of parts of this decision is provided at https://freiheitsfoo.de//files/2013/10/Census-Act.pdf (last visited Oct. 6, 2018).

119 A need "for 'smart transparency' by designing the socio-technical infrastructures" is also referred to by Mittelstadt et al., *supra* note 5, at 10.

120 *Commission Proposal for a Regulation of the European Parliament and of the Council Amending Regulation (EU) No. 168/2013 as Regards the Application of the Euro 5 Step to the Type-Approval of Two- or Three-Wheel Vehicles and Quadricycles*, COM (2018) 137 final (Mar. 19, 2018).

new method may bear unknown risks. [121]This rule can apply to AI *mutatis mutandis*. Accordingly, in cases where AI is ***used*** within medical devices or in the course of surgeries, the functionality and risks associated with the ***use*** of the respective AI needs to be explained in detail. Only after being subject to subject to such **[*207]** detailed information can a patient take an informed decision as to whether or not he or she agrees with the respective medical treatment. Respecting the free will of a person consequently requires at all times that AI technology and its *modus operandi* be understandable and transparent.

One specific case which needs to be considered is autonomous vehicles. In road traffic accidents, the driver and/or owner of a car is typically held liable for eventual damages. This approach, however, may not be appropriate with regard to autonomous cars. Therefore, there is already a debate as to whether the liability in case of accidents caused by autonomous cars should be shifted towards the car manufacturers. [122]

(3) Example 3: Telecommunication Law

On February 17, 2017, the German Federal Network Agency banned a doll called Cayla from being sold and ordered the destruction of all devices which had already been sold. [123]The legal basis of this decision was § 148 (1) no. 2, 90 of the German Telecommunication Act. The rationale was that because of the doll's connectivity to its manufacturer (required because the doll was AIenabled), the doll was effectively a spy on the child, recording all the data the child says to devices including their most precious secrets. [124]Likewise, the agency was concerned that the devices were hackable, exposing children to threats such as pedophilia or ideological communications. Since then, the regulator has ***used*** the law to ban similar devices as well as smart watches. [125]This strict approach adopted to protect children, one of the most vulnerable demographics, has a further legal basis in Art. 16 (1) of the **[*208]** Convention on the Rights of the Child. According to this, "no child shall be subjected to arbitrary or unlawful interference with his or her privacy, family, home or correspondence." [126]

(4) Example 4: Taking Evidence in Court Proceedings

A further example of the possible application of existing regulation to AI is the ***use*** of technical applications for the taking of evidence in court proceedings. Section 244(3) 2 of the German Code of Criminal Procedure holds that an application to take evidence may be rejected if the evidence is wholly inappropriate. On this legal basis, the German Federal Court of Justice has decided expressly that evidence gathered by ***use*** of a certain polygraph-based method

---

121 Bundesgerichtschof [BGH] [Federal Court of Justice] June 13, 2006, case no. VI ZR 323/04, NJW 2006, 2477 (Ger.); confirmed by decision March 27, 2007, case no. VI ZR 55/05, NJW 2007, 2767, 2769 (Ger.).

122 Alexander Hevelke & Julian Nida-Rümelin, Responsibility for Crashes of Autonomous Vehicles: An ***Ethical*** Analysis, 21 SCI ENG ETHICS 619, 620 (2015) https://link.springer.com/content/pdf/10.1007%2Fs11948-014-9565-5.pdf (last visited Oct. 6, 2018);*Commission Staff Working Document on Liability for Emerging Digital Technologies*, 13 SWD 3 (2018) 137 final (Apr. 25, 2018).

123 Press Release, Bundesnetzagentur Removes Children's Doll "Cayla" From the Market, Bundesnetzagentur [BNetzA] [German Federal Network Agency], (Feb. 2, 2017).

124 Kay Firth-Butterfield, *Generation AI: What happens when your child's friend is an AI toy that talks back?*, WORLD ECONOMIC FORUM (May 20, 2018) https://www.weforum.org/agenda/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/. For a legal analysis that was also referred to by the German Federal Network Agency, see Stefan Hessel, "*My friend Cayla*" *- eine nach § 90 TKG verbotene Sendeanlage?*, JurPC Web-Dok. 13/2017, Abs. 1-39, http://www.jurpc.de/jurpc/show?id=20170013 (last visited Oct. 6, 2018).

125 *See, e.g.*, Rebecca Staudenmaier, *Germany bans sale of child-snooping smartwatches*, DEUTSCHE WELLE (Nov. 17, 2017) https://p.dw.com/p/2nqAM (last visited Apr. 13, 2019).

126 United Nations Convention on the Rights of the Child, art. 16 (1), Nov. 20, 1989.

is wholly inappropriate and therefore cannot be relied upon for judicial decision-making purposes. [127]This ruling, which was further confirmed by other German courts, was based on the finding of the court that the specific method for taking evidence by ***using*** polygraphs is not generally and unequivocally accepted among the relevant experts as a correct and reliable method for the taking of evidence. In addition, polygraphs rely on statistical data which cannot be extrapolated to individual cases. Finally, polygraph tests are susceptible to manipulation. [128]

Similarly, AI is susceptible to errors and manipulation. As described above, AI algorithms are based on complex statistical calculations and lack a sufficient degree of transparency so that their mode of operation cannot be entirely understood by humans. [129]It can therefore be concluded from existing German case law that--at least for the time being--AI-driven applications cannot be relied upon for the taking of evidence in court proceedings.

[*209] *b. Eventual Need for New Laws*

(1) Example 1: Defining Red Lines for AI

New regulation by legislation may become necessary to define certain red line areas where AI should not be ***used*** at all or ***used*** only to a strictly limited extent where ***use*** of AI would have disproportionally harmful impacts on individuals or society. [130]While exactly where such red lines should be drawn requires an in-depth debate, three fundamental issues that should be considered for possible legislation are:

> . First, do we want AI-powered humanoid robots with a physical human appearance to become part of our daily lives? It is not necessary for robots to have a physical humanoid appearance. Rather, in order to protect the unique nature and singularity of human life, a corresponding *per se* prohibition could be implemented. Such *per se* prohibition would meet the requirements of a broad and comprehensive protection of human dignity as the most fundamental value [131]under Art. 1 of the Universal Declaration of Human Rights [132]and Art. 1 of the European Charter of Fundamental Rights. [133]The need for a broad protection of the singularity of human life further follows from Art. 3(2) subpara. 4 of the European Charter of Fundamental Rights. According to this, reproductive cloning of human beings is prohibited *per se*. Depending on their technological abilities and physical look, humanoid robots may in the future become more and more confusingly similar to humans, all the more as there is already significant research and **[*210]** development

---

127 Bundesgerichtshof [BGH] [Federal Court of Justice] Dec. 17, 1998, Neue Juristische Wochenschrift [NJW] 657, (658), 1999 (Ger.).

128 Kammergericht [KG] [Higher Regional Court of Berlin] June 2, 2000, case no. 1 AR 573/00 - 4 Ws 110/00, juris; Higher Regional Court of Bremen, May 28, 2001, case no. 5 UF 70/2000b, juris; Federal Court of Justice, June 24, 2003, FPR 2003, 571; Federal Court of Justice, Nov. 30, 2010, NStZ 2011, 474; Federal Administrative Court, July 31, 2014, NVwZ-RR 2014, 887.

129 *See supra* Part I.1.

130 See for instance with regard to facial recognition systems the concerns expressed by Microsoft president Brad Smith. Smith, *supra* note 80.

131 The importance of human dignity as the lens through which to understand and design what a good AI society might look like is also suggested by Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. Cath, et al., *supra* note 8, at 21.

132 Universal Declaration of Human Rights, *supra* note 101. The importance of human dignity is underlined in the preamble, which bases the Declaration upon the "recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family" as "the foundation of freedom, justice and peace in the world."

133 Charter of Fundamental Rights of the European Union, *supra* note 101. Art. 1 expressly reads: "Human dignity is inviolable. It must be respected and protected."

activities ongoing in the field of bioelectronics. [134]Whether from a technical policy-making perspective this could be considered as a new form of reproductive cloning or whether the wording of the Charter should be expanded accordingly, is a question which in view of its fundamental nature needs to be discussed separately.

. Second, stricter legislation may be required in relation to AI-driven weapon systems. [135]In this regard, a group of leading AI scientists has signed a pledge calling " *upon governments and government leaders to create a future with strong international norms, regulations and laws against lethal autonomous weapons*." [136]

. A third critical question is whether AI should be involved in political and judicial decision-making processes. [137]Insofar, a debate is already ongoing as to whether courts should be allowed to ***use*** risk-assessment tools for the purpose of sentencing in criminal cases. [138]

(2) Example 2: Ex Ante Control Requirements for AI Algorithms and Post Launch Market Surveillance

Taking account of the lack of transparency of AI systems and the fact that self-***learning*** algorithms may behave in unexpected ways, "[l]awmakers on national and international levels should be encouraged to consider and carefully review a potential need to introduce new regulation where appropriate, including rules subjecting the market launch of new AI/AS driven technology to **[*211]** prior testing and approval by appropriate national and/or international agencies." [139]The so called "black box concern" therefore, is a major reason why AI algorithms should only be put on the market after prior rigorous testing. [140]A corresponding *ex ante* control regime could either be modelled according to marketing authorization regulations which are, for instance, in place with regard to medicinal products. [141]Alternatively, it could be modelled in accordance with the type approval systems in place for high-tech products such as motor vehicles and parts thereof. [142]

From a legal perspective, the need to establish a marketing authorization or type approval regime for AI applications could in particular be required by the precautionary principle as a guiding policy-making rule.

---

134 Glenn M. Walker, et al., *A Framework for Bioelectronics Discovery and Innovation* 5 (2009). *See, e.g.*, Bozhi Tian, et al., *Macroporous Nanowire Nanoelectronic Scaffolds for Synthetic Tissues*, NATURE MATERIALS, Aug. 26, 2012.

135 IEEE, *supra* note 39, at 113.

136 Future of Life Institute, *Lethal Autonomous Weapons Pledge*, https://futureoflife.org/lethal-autonomous-weapons-pledge/ (last visited Oct. 6, 2018).

137 This was considered by Park in a public expert hearing in front of the German Federal Parliament. Wortprotokoll der 85. Sitzung, Protokoll-Nr. 18/85, Deutscher Bundestag, Ausschuss Digitale Agenda, 22 March 2017, 35, https://www.bundestag.de/blob/526206/65ba7190b0b30f7dbae815d27c8cba80/protokoll-data.pdf (last visited Oct. 6, 2018).

138 *See State v. Loomis: Wisconsin Supreme Court Requires Warning Before **Use** of Algorithmic Risk Assessments in Sentencing* (Mar. 10, 2017).

139 IEEE, *supra* note 39, at 160.

140 Kate Crawford calls for "rigorose Tests, um sicherzugehen, dass sie nicht einseitig oder fehlerhaft sind." Patrick Beuth, *Die Automaten brauchen Aufsicht*, Zeit Online (Oct. 25, 2017), http://www.zeit.de/digital/internet/2017-10/kuenstliche-intelligenz-deepmind-back-box-regulierung (last visited Oct. 6, 2018). IEEE,*supra* note 39, at 7 ("The logic and rules embedded in the system must be available to overseers thereof, if possible, and subject to risk assessments and rigorous testing.").

141 *See supra* note 51. Kate Crawford has recently called for corresponding regulation in the field of AI systems. BEUTH, *supra* note 140.

142 Directive 2007/46/EC of the European Parliament and of the Council of 5 September 2007 Establishing a Framework for the Approval of Motor Vehicles, and Their Trailers, and of Systems, Components and Separate Technical Units Intended for Such Vehicles, 2007 O.J. (L 263) 1.

[143]According to this, "where there is uncertainty as to the existence or extent of risks to human health, the institutions may take protective measures without having to wait until the reality and seriousness of those risks become fully apparent." [144]An emerging view argues that this principle should also be applied to **[*212]** new technologies including AI. [145]Given the complexity of corresponding considerations, this topic, as well, needs to be analyzed in depth separately. In particular, the potential of AI to cause harm needs to be assessed critically. A particular differentiation will have to be made between the development of artificial general intelligence and special purpose AI. For special purpose AI, which can only fulfil a specific limited task such as steering a car or responding to help desk calls, the respective ***use*** case will have a decisive impact on the relevant risk assessment and accordingly on the question whether and to what extent precautionary control measures are necessary. For artificial general intelligence, in view of its potential to behave and define its tasks independently, the more relevant question will be whether and to which extent such technology--should it be possible to be created at some point in time--shall be prohibited completely.

In addition, specific rules on *post* launch market surveillance need to be considered with regard to AI systems in order to avoid unwanted side effects and detrimental developments. This may include the need to set up a specialized administrative agency focusing on the surveillance of AI systems. The application of such rules could be a function of a new type of regulator referred to earlier. [146]

2. International Resolutions and Treaties

Should an analysis of specific ***ethical*** concerns come to the conclusion that binding regulation is necessary but not sufficient to be dealt with on a national level, corresponding issues could also be addressed by international resolutions and treaties of international organizations. As is the case for national legislation, international resolutions and treaties are also binding. However, they are generally only binding upon the parties to the agreement, i.e., the states that signed the corresponding treaties. Public international law, in particular the Vienna Convention on the Law of Treaties, [147]obliges the states to transform international resolutions and treaties into their respective national law. [148]The major advantage of international resolutions and treaties is that they ultimately provide for **[*213]** transnational binding and enforceable rules. In view of the ratification requirement, they are--as is the case for national legislation--subject to a democratic basis ensuring participation of the relevant persons and interest groups.

Some restrictions, however, apply. As a result of the ratification requirement, it is possible for national legislation to transform an international resolution or treaty into national law which provides for national particularities so that the purpose of harmonization is often not achieved. Further, the enforcement of corresponding rules is subject to

---

143 *Communication from the Commission on the Precautionary Principle*, at 8, 10, 13 COM (2000) 1 final (Feb. 2, 2000) (underlining the precautionary principle as a basic rule that aims at protecting consumers against potential harmful developments on the basis of scientific risk assessments). *See also* CJEU, decision of 5 May 1998, C-157/96 and C-180/96, rec. 63 resp. rec. 99 -- *BSE; Neuhäuser ibid.* (fn. 100), p. 284; John Weckert, *In Defense of the Precautionary Principle*, IEEE Technology and Society Magazine, Winter 2012, at 12. It should be noted though that the precautionary principle is still not entirely acknowledged as a governance principle in international law. Didier Bourguignon, *The Precautionary Principle: Definitions, Applications, and Governance*, European Parliamentary Research Service (EPRS) 6 (Dec. 2015).

144 Case C-157/96, The Queen v. Ministry of Agriculture, Fisheries and Food, 1998 ECR I-2211 (1998).

145 Weckert, *supra* note 143, at 12. *But see* Adam Thierer et al., *Artificial Intelligence and Public Policy*, MERCATUS RESEARCH (2017) https://www.mercatus.org/system/files/thierer-artificial-intelligence-policy-mr-mercatus-v1.pdf (last visited Oct. 6, 2018).

146 Scherer, *supra* note 109.

147 Vienna Convention on the Law of Treaties, *opened for signature* May 23, 1969, 1155 U.N.T.S. 331.

148 For Germany, the German Constitution sets out the need for a ratification of international treaties. Grundgesetz [GG] [Basic Law], § 59(2).

national regimes because international law does not provide for immediate international law enforcement. A further downside is that as a consequence of the international law-making process and the often-difficult process of finding compromises between conflicting views of the various states, these resolutions tend to be vague and only provide rough and sometimes unclear guidance. Additionally, the process of finding agreement on an international level is usually extremely long. Irrespectively, very basic and fundamental ***ethical*** principles and values should still be agreed upon on this basis in order to achieve transnational consensus as to the protection of fundamental human values and to underline the singularity and equality of human life. In view of the slow policy-making process on the international level, it remains necessary to consider additional and immediate legal action on a national level to address specific and immediate concerns which may arise from the ***use*** of new technology such as AI.

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. [149]Also, the UN has established its "Centre for Artificial Intelligence and Robotics" in The Hague which shall, amongst other tasks, perform a risk assessment and stakeholder mapping and analysis. [150]More specifically, there is an ongoing debate around an international ban of AI-driven killer robots. [151]

[*214] On July 12, 2018, the UN Secretary General appointed a High-Level Panel on Digital Co-operation. [152]The Secretary General asked the Panel to contribute to the broader public debate on the importance of cooperative and interdisciplinary approaches to ensure a safe and inclusive digital future for all, taking into account relevant human rights norms. [153]In its first report, the Panel made several recommendations to facilitate the development of "a global commitment for digital cooperation." [154]These recommendations included a call for "clear human accountability for autonomous systems." [155]

3. Bilateral Investment Treaties (BITs)

As an alternative to multinational treaties, states could consider addressing AI related concerns in BITs. States could, for instance, agree to establish certain protective measures in relation to AI systems. For example, 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. The benefit in comparison to multinational treaties is that the process of finding an agreement is significantly less complex and that BITs may therefore be put into operation more quickly. Still, such rules are often quite broad and rather vague. Regulation contained in BITs often only provides for indirect protection of ***ethical*** principles. Corresponding agreements can only be ***used*** to enforce

---

149 European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics, 2017 O.J. C 2015/2103; *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*, COM (2018) 237 final (Apr. 25, 2018).

150 *See* UNICRI CENTRE FOR ARTIFICIAL INTELLIGENCE AND ROBOTICS, http://www.unicri.it/in_focus/on/UNICRI_Centre_Artificial_Robotics (last visited Oct. 6, 2018).

151 WEAVER, *supra* note 103, at 142; Toby Walsh, *Why the United Nations Must Move Forward With a Killer Robots Ban*, IEEE (Dec. 15, 2016) http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/united-nations-killer-robots-ban (last visited Oct. 6, 2018).

152 *See Secretary-General Appoints High-Level Panel on Digital Cooperation*, UNITED NATIONS (July 12, 2018) https://www.un.org/press/en/2018/sga1817.doc.htm.

153 *Id.*

154 U.N. Secretary General's High-Level Panel on Digital Cooperation, *The Age of Digital Interdependence* (June 2019).

155 *Id.*

national protective legislation against foreign companies that have their registered seat in a state with whom a BIT is in place without risking investment treaty arbitration proceedings. For instance, if country X, where national legislation allows for unlimited ***use*** of AI in medical devices, enters into a BIT with country Y, where national legislation makes the ***use*** of AI in medical devices subject to an ex-ante marketing authorization requirement, a free trade agreement would usually provide that products from both countries can be sold freely on the respective two markets. As this would obviously put medical device companies residing in country Y at a disadvantage, country Y would usually require the implementation of a similarly strict marketing authorization requirement in country X or would negotiate an exemption from the free trade provisions with regard to medical devices. In case country X agreed to implement a similar marketing **[*215]** authorization requirement, the question whether the appropriate level of protection is indeed provided for in country X would, in case of a dispute, have to be decided within investment treaty arbitration proceedings aimed at amending the legislation in country X. Therefore, the conclusion of BITs in practice often requires the parties to the agreement to adapt their respective legal regimes. For instance, Korea, when entering into a free-trade agreement with the EU, established a more transparent regulatory system in the field of pharmaceuticals. [156]

4. Self-Regulation and Standardization

An industry-driven private regulation approach can address the territorial limitation of state laws as well as the procedural complexity and length of legislative processes. With regard to AI and autonomous systems, technology standards could be developed that make ***use*** of technical measures providing for ethically compliant behavior by AI algorithms. That includes privacy by design, transparency by design, as well as potential kill switch technologies.

The benefit of self-regulation is that such approach is driven by industry and technology specific experts. The territorial applicability of technology standards would not be limited in scope. At the same time, a plurality of opinions and ***ethical*** regimes could be maintained. Potential disadvantages of technology standards are that their development may lack democratic legitimization and participation of the public. [157]Since technology standards are generally agreed upon between industry stakeholders--which often involving competing companies--it is further crucial to comply with applicable competition law requirements. [158]From a competition law perspective, to what extent technology standards employing ***ethical*** principles for AI systems can be agreed upon is a question that needs to be carefully assessed and answered on a case-by-case basis.

For example, technology standards such as transparency and identity tag standards could address accountability issues and ensure that AI systems record decisions taken and considerations relied upon by the AI. Further, to ensure controllability of AI and autonomous systems, one could consider the implementation of "kill-switch" **[*216]** technology and agree on corresponding standards. Ultimately, one could even think of guardian AI systems that aim to ensure compliance of AI and autonomous systems with legal or other regulatory preconditions.

5. Certification

Similar to technology standards, compliance with ***ethical*** principles can be achieved by establishing certification systems. Certification systems offer the general advantages and legal concerns of self-regulation-based governance approaches as referred to above in relation to technical standardization. However, certification

---

156 European Commission, Das Freihandelsabkommen zwischen der EU und Korea in der Praxis, (2011), http://trade.ec.europa.eu/doclib/docs/2011/october/tradoc_148307.pdf.

157 Tim Büthe & Walter Mattli, THE NEW GLOBAL RULERS: THE PRIVATISATION OF REGULATION IN THE WORLD ECONOMY, 220 (2011).

158 In Europe, Art. 101 TFEU has to be complied with. The European Commission has explained its rather generous approach as to the competition law assessment of standardization agreements. *See* Guidelines on the applicability of Article 101 of the Treaty on the Functioning of the European Union to horizontal co-operation agreements, 2011 O.J. (C 11) 1.

organizations may have the benefit of being exempt from application of competition laws. To what extent and under which circumstances that is the case, is subject to an ongoing discussion. [159]

6. Contractual Rules

As an alternative to collective self-regulation measures such as standardization and certification, companies can opt to comply with certain ***ethical*** values and principles on a contractual basis ***using*** bilateral agreements. This is standard business practice; an example is a manufacturer-supplier relationship where so-called compliance clauses are implemented to make sure that no products made by exploitation of child labor are supplied. Respective contract clauses could be extended to the obligation of the parties to only ***use*** AI systems which comply with specific ***ethical*** principles.

A contractual approach is probably the most flexible way to ensure ***ethical*** compliance. Also, enforcement is relatively efficient and may be sought through the civil court system and alternative dispute resolution means. The disadvantage is that corresponding ***ethical*** rules would only be binding upon the parties to the contract.

[*217] *7. Soft Law*

Finally, 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. Soft law is consequently more flexible and can be adjusted to technical developments more easily. Also, soft law can be more specific than binding laws and can go--at least to a certain extent--beyond the usual minimum consensus which legislation by national and international organizations typically only manage to agree upon. The obvious downside is that soft law is not binding and not enforceable.

8. Agile Governance

Given the difficulties enumerated above and recognizing that AI implementing technologies are developing so swiftly that it is almost impossible for traditional legislation to keep up with them let alone get ahead, the World Economic Forum has created an 'agile governance' approach which incorporates many of the ideas in this white paper. [160]The basic observation is that governments are ***responsible*** for protecting citizens from various harms caused by new products and technologies; this is traditionally accomplished by holding perpetrators accountable once the harm has occurred. With AI impacting society at unprecedented speed, scope, and scale, governments must protect the public before the harm occurs by promoting the ***responsible*** design, development and ***use*** of this transformative technology. This requires a more agile type of regulator (i.e., one that is proactively working with companies to ensure safety up front and not after-the-fact), without stifling the many societally beneficial ***uses*** of AI. [161]The regulator of the future must be expert, nimble and work with companies to certify their products as fit for their purpose. This will not only protect citizens but also encourage innovation in the AI space because companies will not be at risk of wasting R&D expenditures on products that may be banned or regulated in the future.

---

[159] Landgericht Köln [LG] [Regional Court of Cologne] Mar. 12, 2008, RECHTSPRECHUNG DER OBERLANDESGERICHTE IN ZIVILSACHEN [OLGZ] 1, 2008 (Ger.) (denying the applicability of competition law on the grounds that such organization is not acting commercially); Oberlandesgericht Düsseldorf [OLG] [Higher Regional Court of Düsseldorf] Mar. 30, 2011, RECHTSPRECHUNG DER OBERLANDESGERICHTE IN ZIVILSACHEN [OLGZ] 1, 2011 (Ger.) (tending towards applicability of competition law but actually referring the question to the CJEU); Oberlandesgericht Düsseldorf [OLG] [Higher Regional Court of Düsseldorf] Aug. 14, 2013, RECHTSPRECHUNG DER OBERLANDESGERICHTE IN ZIVILSACHEN [OLG] 1, 2013 (Ger.) (questioning the applicability of competition law ultimately left open by CJEU).

[160] *Agile Governance: Reimagining Policymaking in the Fourth Industrial Revolution*, WORLD ECONOMIC FORUM, (Apr. 2018), https://www.weforum.org/whitepapers/agile-governance-reimagining-policy-making-in-the-fourth-industrial-revolution.

[161] The need for more flexibility on the side of regulators is also acknowledged by European Parliamentary Research Service. *See supra* note 6, at 17 (discussing logistics and transport as a ***use*** case for new digital technologies).

[*218] *C. Monetary Incentives*

In addition to the adoption of specific policy-making instruments, regulators have the option to create monetary incentives to guide the development and implementation of new technologies in line with certain policy goals. With regard to AI applications, regulators could, for example, subject the grant of research and development funding to the condition that respective R&D proposals and their results will comply with specific ***ethical*** requirements. To this end, the relevant core ***ethical*** principles would need to be defined as the first step, for instance within the framework of an ethics charter for AI applications. [162]Such ethics charters could be issued in the form binding legislation or as soft law. Second, reference to mandatory compliance with such ***ethical*** principles would need to be made in research and development grants.

In view of the currently envisaged extensive amounts of funding to be granted for the benefit of research and development projects in the field of AI, [163]it appears that concrete steps should be initiated immediately in order to ensure ethics compliance by AI applications. Conditioning research and development project funding with specific ***ethical*** requirements would ensure that, from the very beginning, companies would only develop and market such technology and related business models as are in line with the core values of our society.

III. TWO PRACTICAL APPROACHES TO IMPLEMENTING 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 for the implementation of ethics into AI and autonomous systems.

[*219] *A. The IEEE Global Initiative*

The IEEE Global Initiative is a program of the Institute of Electrical and Electronics Engineers ("IEEE") launched in December 2015. A primary goal of the IEEE Global 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. [164]To this end, the IEEE Global Initiative issued a document titled "Ethically Aligned Design -- A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems." [165]This describes the so-called IEEE P7000 TM series specific proposals for actual operational standards which can be adopted by designers of AI and autonomous systems. [166]

The report "Ethically Aligned Design" summarizes insights and recommendations that provide a key reference for the work of technologists in the related fields of science and technology who are developing and programming AI and autonomous systems. The document first identifies pertinent "Issues" and "Candidate Recommendations"

---

[162] Consider in particular the work done by the European Group on Ethics in Science and New Technologies. *See infra* Part IV.2.a. The European Commission has issued a communication stating that "AI ethics guidelines" should be developed by the end of the year 2018. EUROPEAN COMMISSION, "Communication from the Commission to the European Parliament, the European Council, the European Economic and Social Committee and the Committee of the Regions -- Artificial Intelligence for Europe", SWD (2018) 137 final, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0237&from=EN.

[163] *See* European Commission, *supra* note 162.

[164] Raja Chatila et al., *IEEE Global Initiative Aims to Advance* ***Ethical*** *Design of AI and Autonomous Systems*, IEEE SPECTRUM, (Mar. 2017), https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/ieee-global-initiative-***ethical***-design-ai-and-autonomous-systems.

[165] IEEE, *supra* note 44. The first version of the document also provides useful insights. *See Ethically Aligned Design*, IEEE (Dec. 2016), http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf (last visited Oct. 6, 2018).

[166] Standardization projects of the P7000 series exist. *See* IEEE, *supra* note 44 at 4; ETHICS IN ACTION, https://ethicsinaction.ieee.org/ (last visited Oct. 6, 2018).

which facilitate the emergence of national and global policies that align with these principles. [167]Next, the document, and in particular its "Candidate Recommendations," can be ***used*** as a basis for the development of operational standards. [168]

[*220] One of the key concerns over AI and autonomous systems is that their operations must be sufficiently transparent for users, authorities and courts. [169]The IEEE P7001 TM standard intends to provide a guide for designers for self-assessing transparency during development and suggests mechanisms for improving transparency. This includes, for instance, the need for secure storage of sensor and internal state data, comparable to a flight data recorder. [170]A further major concern relates to the maintenance of privacy. [171]The IEEE Global Initiative addresses this concern in particular with its standardization proposal IEEE P7002 TM. The purpose of providing a standardized "Data Privacy Process" (IEEE P7002 TM) is to manage ***ethical*** issues for systems and software that collect personal data. The standard defines requirements for corporate data-collection policies and quality assurance. It includes a ***use*** case and data model for organizations developing applications. The standard also provides designers with ways to identify and measure privacy controls in their systems. [172]

B. World Economic Forum's Project on Artificial Intelligence and ***Machine Learning***

The World Economic Forum, with its focus on international public-private partnerships, is building an excellent neutral and objective platform to help countries as well as businesses struggling with policy implementation and governance of AI. It has a number of projects on AI governance as well as other projects on governance of drones, blockchain, autonomous vehicles, the environment and technology, IoT, precision medicine, cross-border data flows, and ecommerce. All projects are required to include ethics and values, social inclusion and human centered design. The Forum is establishing Centers for the Fourth Industrial Revolution in San Francisco, Tokyo, Beijing and

---

[167] IEEE, *supra* note 44 at 3.

[168] Concrete standardization proposals that are currently being discussed and developed are the following:

IEEE P7000 TM -- Model Process for Addressing ***Ethical*** Concerns During System Design IEEE P7001 TM -- Transparency of Autonomous Systems

IEEE P7002 TM -- Data Privacy Process

IEEE P7003 TM -- Algorithmic Bias Considerations

IEEE P7004 TM -- Standard on Child and Student Data Governance

IEEE P7005 TM -- Standard for Transparent Employer Data Governance

IEEE P7006 TM -- Standard for Personal Data Artificial Intelligence (AI) Agent

IEEE P7007 TM -- Ontological Standard for Ethically Driven Robotics and Automation Systems

IEEE P7008 TM -- Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems

IEEE P7009 TM -- Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

IEEE P7010 TM -- Wellbeing Metrics Standard for ***Ethical*** Artificial Intelligence and Autonomous Systems

[169] *See supra* Part I.3.c.

[170] IEEE P7001 TM.

[171] See *supra* Part I.3.e.

[172] Monica Rozenfield, *Seven IEEE Standards Projects Provide* ***Ethical*** *Guidance for New Technologies*, THE INSTITUTE, (May, 2017), http://theinstitute.ieee.org/resources/standards/seven-ieee-standards-projects-provide-***ethical***-guidance-for-new-technologies.

Mumbai. It will also establish 'Affiliate Centers' globally. At these Centers, governance projects for AI and other technologies will be co-created with governments, businesses, academics and civil society. Currently, the following projects are ongoing:

[*221] *1. Unlocking Public Sector AI*

Although AI holds potential for vastly improving government operations, many public institutions are cautious about harnessing it because of concerns over bias, privacy, accountability, transparency and overall complexity. Baseline standards for effective and ***responsible*** procurement and deployment of AI by the public sector 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.

2. 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 the concerns about the need to design, develop and deploy it responsibly. A practical set of tools can assist Board Members and decision-makers in asking the right questions, understanding key trade-offs, and meeting the needs of diverse stakeholders, as well as how to consider and optimize approaches, such as appointing a Chief Values Officer or creating an Ethics Advisory Board.

3. Generation AI

This project specifically deals with the development of standards for protecting 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, special 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 have a part in understanding the design and values of these algorithms, and prevent biases in AI training data and algorithms from undermining educational objectives. Transparency and accountability can build the trust necessary to accelerate the positive social benefits of these technologies for all. [173]

4. Teaching AI Ethics

Decisions regarding the ***responsible*** design of AI are often made by engineers who receive little training in the complex ***ethical*** considerations of 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 **[*222]** Future Councils 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. [174]

5. The Regulator of the Future

Another way of addressing the problem of adequate implementation of ethics into AI is to re-imagine the regulator to ensure that citizens, companies and governments are all capable of understanding and ***using*** advanced technologies while at the same time able to develop appropriate and risk-aware policies through a collaborative process. This is work being undertaken by the World Economic Forum out of its office in San Francisco. [175]

---

[173] For a practical example, see the Cayla decision of the German Federal Network Agency, in Part II.2.a.cc.(1)(c).

[174] *See Artificial Intelligence and Robotics*, WORLD ECONOMIC FORUM: STRATEGIC INTELLIGENCE, https://intelligence.weforum.org/topics/a1Gb0000000pTDREA2?tab=publications (last visited Oct. 14, 2019).

[175] *See generally*, CENTRE FOR THE FOURTH INDUSTRIAL REVOLUTION, https://www.weforum.org/centre-for-the-fourth-industrial-revolution; *supra* Part II.2.h. For details see the WEF White Paper "Agile Governance Reimagining Policy-

IV. AI GOVERNANCE: DETERMINING THE APPROPRIATE REGULATION DESIGN

The previous sections have shown that policy-makers can choose from a broad range of regulatory measures that enable them to determine a fine-tuned AI governance regime taking into account the particularities and possible impacts of AI and autonomous systems. Designing an appropriate AI governance regime requires an in-depth assessment of whether any regulation exists that can deal with the challenges associated with the increasing ***use*** of AI and autonomous systems adequately, [176]or whether new regulation is needed. If regulation is required, the next question from a policy-making perspective is which regulatory instrument should be chosen. Answering these questions is a complex challenge as a careful risk assessment--often referred to as "impact assessment"--has to be conducted. This assessment is particularly complex with regard to the issue of AI-governance.

[*223] *A. The Need to Conduct Risk Assessments for New Technologies*

New technologies are generally driven by optimistic expectations of potential benefits which researchers and developers intend to achieve. However, new technologies, at the same time, always entail new risks. This can be illustrated best by examining the advent of nuclear power. The optimistic expectation was that this new technology would resolve the world's energy supply problem. The consequences were the development of nuclear weapons and the fact that there is no environmentally friendly and sustainable way to deal with the nuclear waste. So what lesson is to be ***learned***? Should society abstain from new technologies in view of the 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 in general?

In a pragmatic sense, this question can only be answered in the negative. Because AI is already being ***used*** and developed, we need to focus on how to make good ***use*** of AI and how to avoid irreparable harm. History's lessons should tell us to be cautious and assess potential risk scenarios carefully before implementing and establishing a potentially risky and uncontrollable new technology. [177]On this basis, abuse and risk prevention means and mechanisms should be employed. A corresponding risk assessment and scientific review involving relevant experts and persons concerned may even result in the definition of ***use*** cases where a certain technology like AI should not be employed at all. [178]For other ***use*** cases, specific preconditions such as the need to pursue marketing authorization procedures or implement specific security technologies may have to be considered. [179]

Obviously, this may result in the need for additional regulation and corresponding law enforcement actions. However, this process, and the regulations which may ultimately be found to be appropriate as a consequence of such risk assessment, should be considered as a necessary precaution before moving forward into a more digitalized and automated living and working environment in order to avoid opening another Pandora's box. Finally, the conduct of risk-benefit assessments and consequential implementation of risk and abuse prevention mechanisms not only protects people and their **[*224]** fundamental rights but further increases general acceptance of new technologies and, therefore, ultimately results in economic welfare gains.

B. The Complexity of AI-Governance

---

making in the Fourth Industrial Revolution" at http://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf (last visited Apr. 13, 2019).

176 *See* Microsoft, THE FUTURE COMPUTED -- ARTIFICIAL INTELLIGENCE AND ITS ROLE IN SOCIETY 78 (2018).

177 *Cf.* René *von Schomberg, From Ethics of Technology towards an Ethics of Knowledge Policy & Knowledge Assessment*, THE EUROPEAN COMMISSION 15 (Jan. 1, 2007), https://publications.europa.eu/en/publication-detail/-/publication/aa44eb61-5be2-43d6-b528-07688fb5bd5a (last visited April 13, 2019).

178 *See supra* Part II.2.a.cc.(2)(a).

179 *See supra* Part II.2.a.cc.(2)(b).

Designing an appropriate AI-governance system is particularly difficult for several reasons. First, because of the diverse nature of ***ethical*** concerns. Second, due to the difficulty of determining the appropriate regulatory instrument. Third, because there are complex interactions between the relevant technology, the economy and markets, individual humans and the society as well as the environment and, ultimately, politics and regulation.

*1. **Ethical** Diversity*

The political debate is now addressing the urgent topic of the ***ethical*** and societal implications which the digital transformation in general, and AI in particular, is likely to have. A comprehensive list of ***ethical*** concerns has been presented by the European Commission's Group on Ethics in Science and New Technologies. These are summarized in the table below: [180]

The *ethical* principles of the European Group on Ethics in Science and New Technologies		
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

*Fig. 1 -- The **ethical** principles of the European Group on Ethics in Science and New Technologies*

As indicated above, it is not the purpose of this Article to discuss the content related to details around ***ethical*** principles that might be incorporated by AI applications. This requires a separate, broader and fundamental debate across national, religious and cultural boundaries. What is particularly relevant for the topic dealt with herein is the variety and diversity of ***ethical*** values, their priorities and relationship between them.

[*225] There are fundamental and universal concerns. For instance, Art. 2 of the Treaty on European Union states:

> 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. [181]

Fundamental human values are further set out in the UN's Universal Declaration of Human Rights and other declarations, which expand on these rights for specific groups, such as children. [182] In contrast, other ***ethical*** concerns reflecting specific beliefs of certain individual ***ethical*** 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 when it comes to the possible regulation of ethics. Even a fundamental and generally accepted need--for example the protection of human dignity--may be controversial when defining specific requirements and duties to be complied with by concrete AI applications.

---

180 For details, see European Group on Ethics in Science and New Technology (EGE), *supra* note 10, at 16.

181 Consolidated Version of the Treaty on European Union, 2012 O.J. (C326).

182 Universal Declaration of Human Rights, *supra* note 101; Convention on the Rights of the Child, *supra* note 126; G.A. Res. 44/25, 44 U.N. GAOR Supp. No 49, U.N. Doc. A/44/736 (1989).

An assessment of ***ethical*** implications of AI applications also strongly depends on the relevant cultural and economic framework conditions. This is particularly apparent in the field of education and is addressed in the work of the World Economic Forum Teaching ***Ethical*** AI project. For example, from a US and European perspective the Cayla decision of the German Federal Network Agency [183] will generally be considered to be 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 which arise around the ethics of AI are writ large in this microcosm. Privacy, bias, surveillance, manipulation, democracy, transparency and accountability can all be challenged with an AI-enabled toy.

However, an ***ethical*** evaluation may be different from the perspective of developing countries. Most economists believe that accelerating and increasing access to education in developing countries is the best way to close the gap between the developed and developing world. [184] The difficult question to be answered by **[*226]** regulators, then, is how these possible benefits can be balanced with the additional burdens and tasks to be borne on the side of the relevant AI companies. For instance, if a regulator should infer that an AI company may have access to children's data through AIenabled toys offered for educational purposes, should there be a duty on the company to red flag children who share suicidal thoughts, other self-harming behavior or threat scenarios? Ethically, one could argue that technology enables a company to protect a child's life by informing its parents of possible dangerous scenarios. Whether privacy and private autonomy or the protection of a child's health and life should have greater weight, however, will most likely not be decided unanimously across the globe.

2. Selecting the Appropriate Regulatory Instrument

Good AI governance requires the right regulatory instrument to be chosen for each ***ethical*** concern. 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 corresponding graded governance model can be illustrated as follows:

Fig. 2 -- Graded governance model for the implementation of ***ethical*** concerns in AI systems

In view of the diversity of ***ethical*** values explained above, it must be acknowledged that there can be no "one size fits all" solution. As has been pointed out before, formal legislation may in particular be required under principles such as the German **[*227]** constitutional principle of " *Parlamentsvorbehalt*" in case the ***use*** of new technologies has material impacts on the protection of fundamental rights and constitutional principles. [185] Also, the obligation not to cause harm to other people, the need to compensate with damages in case harm is caused and the obligations to respect personal rights, autonomy and privacy are generally subject to regulation by statutory laws on national and international level. In this regard, the precautionary principle may further call for binding legislation. [186]

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

---

183 *See supra* Part II.2.a.cc.(1)(c).

184 Børge Brende, *Why education is the key to development*, WORLD ECONOMIC FORUM (July 7, 2015), https://www.weforum.org/agenda/2015/07/why-education-is-the-key-to-development/ (last visited Apr. 13, 2019).

185 *See supra* Part II.2.a.bb.

186 *See supra* Part II.2.a.cc.(2).

In addition to the various policy-making instruments explained above, development of technological standards that provide for technical solutions complying with specific regulatory requirements should be considered. For care robots, the employment of an AI design that respects the user's will as its guiding principle for its operation could be made by compliance with a respective technology standard while a different standard could be developed for a more paternalistic AI system design. Which kind of technology standard is employed could be indicated to users by the reference to a certain certificate. As indicated above, regulators should, in addition, consider the grant of specific monetary incentives to ensure the compliance of AI applications with ***ethical*** requirements. Because AI is an emerging new technology, it appears to be a particularly effective to subject the grant of research and development funding to compliance with specific ***ethical*** principles. [187]

3. The Magic Square of Regulation in Technological Societies

The third reason why AI-governance is a particularly complex and difficult task is that all relevant parameters are directly or at least indirectly interrelated with each other. The increasing ***use*** of AI and autonomous systems has a direct impact on humans, society and the environment. Existing jobs may become obsolete, **[*228]** new jobs arise, there is less social interaction and more man-to-***machine*** communication and more raw materials may be consumed for the increasing construction of ***machines***. [188]At the same time, new technologies call for new business opportunities and thereby can shape new markets or re-shape existing markets. Depending on the nature of the impacts of these new technologies, politics and the state may be called to consider new regulatory actions. Regulation, however, implies a value decision which needs to be made in light of various, sometimes even contradicting, fundamental principles. This includes the principle of competition as a supposed 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 any action or reaction by one of the aforementioned stakeholders can immediately impact the other aspects and stakeholders. Also, regulation can again impact innovation dynamics. However, regulation may foster the development of new technologies and technology-focused business models. As mentioned already, an example referred to above is data privacy regulation, which on the one side restricts the free ***use*** of personal data but at the same time incentivizes businesses to develop privacy-by-design solutions and thereby contributes to a high level of data protection. [189]All decisive factors including technology and innovation, politics and state, humans, society, environment, as well as the economy and markets are directly interrelated with each other. Whether new technologies require new or amended regulation needs to be decided in light of this complex reciprocal interdependence taking into account normative considerations regarding fundamental rights, constitutional principles, ethics, and competition theories. This relationship between the affected stakeholders and the principles to be referred to for regulation purposes can, therefore, be best described as a magic square, which is illustrated as follows:

[*229] Fig. 3 -- Magic Square of Regulation in technology-driven societies

Finding the right solution for regulation within this magic square in view of new digital and AI driven technologies is a particular challenge because the technology changes rapidly and we cannot guess where the technology will be in five years. In addition, innovation cycles are typically extremely short in the field of digital technologies including AI and autonomous systems so that regulation is challenging in this field.

C. The Question of When to Regulate

In view of the increasingly shorter innovation cycles, policy-makers also need to deal with the question of when to regulate. Overhasty regulatory actions need to be avoided in order to provide for efficient and effective protection of

---

187 *See supra* Part II.3.

188 For the various concerns associated with the increasing ***use*** of AI and autonomous systems, see *supra* Part I.3.

189 *See supra* Part II.2.a.aa.

fundamental rights. At the same time, policy-makers need to make sure that necessary regulation is implemented sufficiently early to avoid new technologies causing irreparable harm. One need only think of the hypothetical situation which humanity would face if there had been forethought of the possible dangers associated with the ***use*** of nuclear energy. Had humanity foreseen the considerable nuclear waste created by nuclear power it would have regulated smarter and consequently developed smarter technologies from the beginning. This example should illustrate that thinking of possible dangers and ways to address and avoid these should be the first step before implementing new technologies, particularly in cases such as AI where operating modes and impacts cannot be entirely foreseen. Now is the time to carefully evaluate possible risks and consider ways to **[*230]** exclude, or at least limit, such risks. In particular, we should consider the precise definition of certain red lines for AI [190] and consider whether, in view of a sensible application of the precautionary principle, AI algorithms, at least with regard to certain ***use*** cases, should be subjected to an appropriate control system. [191]

SUMMARY AND OUTLOOK

The increasing ***use*** of AI and autonomous systems will have revolutionary impacts on society. Despite many benefits, AI and autonomous systems involve considerable risks that need to be managed. Minimizing these risks will emphasize the respective benefits while at the same time protecting the ***ethical*** values defined by fundamental rights and basic constitutional principles, thereby preserving a human centric society. This Article advocates for the need to conduct in-depth risk-benefit-assessments with regard to the ***use*** of AI and autonomous systems.

This Article points out major concerns in relation to AI and autonomous systems such as possible job losses, causation of damages, lack of transparency, increasing loss of humanity in social relationships, loss of privacy and personal autonomy, potential information biases and the error proneness, and susceptibility to manipulation of AI and autonomous systems. This critical analysis aims to raise awareness on the side of policy-makers to sufficiently address these concerns and design an appropriate AI governance regime with a focus on the preservation of a human-centric society. Raising awareness for eventual risks and concerns should not be misunderstood as an anti-innovative approach. Rather, it is necessary to consider risks and concerns adequately and sufficiently in order to make sure that new technologies such as AI and autonomous systems are constructed and operate in a way that is acceptable for individual users and society as a whole. It is of utmost importance to design a sufficiently protective, forward-thinking and visionary AI governance regime that in addition to potential benefits considers the relevant risks in order to make sure that AI and autonomous systems can be ***used*** in an effective and adequate manner to the benefit of humanity.

As a basis for the design of a corresponding visionary AI governance regime, this Article further outlines the various possible policy-making instruments. The variety of such instruments, which policy-makers can make ***use*** of, underlines that ***ethical*** concerns do not necessarily need to be addressed by legislation or international conventions. Depending on the ***ethical*** concern at hand, alternative **[*231]** regulatory measures such as technical standardization or certification may be preferable. For individual ***ethical*** concerns, even bilateral contractual agreements may be sufficient. As suggested herein, an approach to develop a corresponding 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. With regard to the latter, regulators should in particular take their own responsibility seriously and only support research and development compliant with fundamental ***ethical*** principles and values.

In view of the AI's potential revolutionary impact, it is of utmost importance to further raise awareness for the need to consider ***ethical*** considerations not only on the side of policy-makers but also on the side of companies and designers of AI and autonomous systems. The IEEE Global Initiative and the World Economic Forum's projects are the first concrete global approaches. From a legal perspective, more projects should be pursued by additional

---

190 *See supra* Part II.2.a.cc.(2)(a).

191 *See supra* Part II.2.a.cc.(2)(b); note in particular the suggestion made by Smith, *supra* note 80.

stakeholders, because ***ethical*** concerns are highly diverse in nature. Maintaining ***ethical*** diversity is an ***ethical*** concern of its own as this ensures the protection of individuality as a core human value. ***Ethical*** diversity can, however, only be maintained if policy-makers promote the establishment of different solutions which meet the varied concerns of diverse stakeholders and institutions. At the same time, fundamental and universal ***ethical*** values need to be addressed on an international and cross-cultural basis.

Businesses should bear in mind that ensuring ethics compliance for their AI applications will ultimately turn out to be a strong competitive advantage. Ethically aligned products will ultimately be more acceptable to customers. With regard to privacy as one of the core concerns associated with the increasing ***use*** of AI, the European Political Strategy Centre expressly pointed out that "by respecting the legitimate right to privacy of users, AI technologies would be more readily accepted by society at large." [192]This underlines that beyond building a human-centric AI society, due consideration of ***ethical*** concerns can turn into an immediate competitive advantage. Regulators and businesses should therefore share a common interest in ensuring that AI and autonomous systems provide a strict and high level of protection of ***ethical*** values.

Duke Law & Technology Review

End of Document

---

192 EPSC Strategic Notes, *supra* note 6, at 6.