Silja Vöneky

Key Elements of Responsible Artificial Intelligence – Disruptive Technologies, Dynamic Law[1]

"We're making tools not colleagues,
and the great danger is not appreciating the difference,
which we should strive to accentuate,
marking and defending it with political and legal innovations.
(…) We don't need artificial conscious agents.
(…) We need intelligent tools."[2]
Daniel C. Dennett

"We may hope that machines will eventually compete with men in all purely intellectual fields."[3]
Alan M. Turing

One major challenge of the 21st century to humankind is the widespread use of Artificial Intelligence (AI). Hardly any day passes without news about the disruptive force of AI – both good and bad. Some warn that AI could be the worst event in the history of our civilization. Others stress the chances of AI diagnosing, for instance, cancer, or supporting humans in the form of autonomous cars. But because AI is so disruptive the call for its regulation is wide-spread, including the call by some actors for international treaties banning, for instance, socalled "killer robots". Nevertheless, until now there is no consensus how and to which extent we should regulate AI. This paper examines whether we can identify key elements of responsible AI, spells out what exists as part "top down" regulation, and how new guidelines, such as the 2019 OECD Recommendations on AI can be part of a solution to regultate AI systems. In the end, a solution shall be proposed that is coherent with international human rights to frame the challenges posed by AI that lie ahead of us without undermining science and innovation; reasons are given why and how a human rights based approach to responsible AI should inspire a new declaration at the international level.

Introduction

Everything about AI is a hype. It is labeled a disruptive technology. Its transformative force is compared to that of electricity. It is said that just as electricity transformed peoples' lives and industries 100 years ago, AI will now transform our lives.[4] As we are incorporating AI systems into our life, we benefit from the efficiencies that come from AI systems (AIs).[5]

However, a technology like AI is, first of all, a tool. I argue, as the philosopher *Daniel C. Dennett* argues, that AIs are tools and should be regarded and treated as tools. They are tools with a specific quality and power, because AI systems can be used for multiple purposes, and will imitate and replace human beings in many intelligent activities, shape human behavior and even change us as human beings in the process[6] in intended and unintended ways.[7] But even if AIs could be in principle as autonomous as a person they lack our vulnerability and mortality.[8]

This means that as long as we develop, sell and use AI, we can and have to decide how we frame the rules and norms governing AI. As always when we have the chance to get a new, powerful technological tool, we have to answer the question how we can make sure that we as a society will make the right choices – or at least minimize the risk that we will make the wrong choices; and how do we decide what is right and wrong – especially as the field of AI is an area hardly anybody understands fully. I argue that these are questions that cannot be answered

1 The background of this paper is my research on questions of democratic legitimacy in ethical decision making as a Director of an Independent Max Planck Research School in Heidelberg on biotechnology governance, and on the governance of existential risks as a Fellow at Harvard Law School (2015–2016). I am grateful for the inspiration and exchange with the members of our FRIAS Saltus Research Group "Responsible AI", *Philipp Kellmeyer* (Neurology, Neuroethics), *Oliver Müller* (Philosophy), and *Wolfram Burgard* (Robotics) over the last months. I want to thank the research assistants *Tobias Crone, Isabella Beck, Eva Böning*, and *Gideon Wheeler* for their valuable support.

2 *Daniel C. Dennett*, What can we do?, in John Brockman (ed.), Possible Minds – 25 Ways of Looking at AI, 2019, 46, 51.

3 *Alan M. Turing*, Computing Machinery and Intelligence, Mind LIX, 1950, 433 et seq. (reprinted in *Margaret A. Boden* (ed.), The Philosophy of Artificial Intelligence, 1990, 65 et seq.).

4 *Andrew Ng*, in *Martin Ford* (ed.), Architects of Intelligence, 2018, 185, 190.

5 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 477, 484.

6 *Norbert Wiener*, The Human Use of Human Beings, 1954, 96.

7 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 478; *Daniel C. Dennett*, What can we do?, in *John Brockman* (ed.), Possible Minds – 25 Ways of Looking at AI, 2019, 43.

8 *Daniel C. Dennett*, ibid., 51 et seq.

by individuals, corporations or States, only, but have to be answered by the international community as a whole, as well, because AI research, development and deployment, and the related effects are not limited to the territory of a State but are transnational and global.

This paper is a starting point to discuss key elements of responsible AI. Although the notion of intelligence in Artificial *Intelligence* might suggest otherwise, AI as a technology is not per se "good", neither is it "bad". The *first part* spells out features of AI systems, and identifies benefits and risks developing and using AI systems, in order to show challenges for regulating these tools (*see below I*).

The international governance dimension is stressed in the *second part*. There I will look closer at the Recommendations on Artificial Intelligence by the Organisation for Economic Co-operation and Development (OECD) that were adopted in 2019 (*see below II*).[9] These are the first universal international soft law rules that try to govern and frame AI in a general way.

Thirdly, I argue that we should stress the link between human rights and the regulation of AI systems, and highlight the advantages of an approach in regulating AI that is based on legally binding human rights that are part of the existing international legal order (*see below III*).

I. AI Systems as Multipurpose Tools – Challenges for Regulation

1. Notions and Foundations

When we try to understand what AI means as a technology, we realize that there seem to be many aspects and applications relevant and linked to AI systems: from facial recognition systems, to predictive policing, from AI called AlphaGo playing the game GO, to social bots and algorithmic traders, from autonomous cars to – maybe even – autonomous weapons.

A first question we should answer is: How can we explain AI to someone who does not know what AI is, but wants to join and should join the discourse on regulation and governance? A simple start would be to claim, that a key feature of the field of AI is the goal to build intelligent entities.[10] An AI system could be defined as a system that is intelligent, i.e. rational, in the way and to the extent that it does the "right thing", given what it knows.[11] However this is only one definition of an AI system. The standard textbook quotes eight definitions by different authors laid out along two dimensions including two aspects to measure the success of an AI system in relation to human performance ("thinking humanly"; "acting humanly"); and two aspects to measure the success of an AI system in relation to ideal performance ("thinking rationally"; "acting rationally").[12] But even if those are correct who state that AI is concerned with rational or intelligent behavior in artifacts, the underlying question is whether it is correct to state that the notion of "intelligence" means the same as the notion of "rationality".[13] It seems reasonable to claim that AI systems exhibit forms of intelligence that are qualitatively different to those seen in humans or animals as biological agents.[14]

As a basic description one might state that AI tools are based on complex or simple algorithms[15] used to make decisions, and are created to solve particular tasks. Autonomous cars, for instance, must drive (in a given time without causing accidents or violating laws) to a certain place, and game-playing AI systems should challenge or even win against a human being.[16]

As AI is expected to fulfill a certain task, there are required preconditions for a system to be able to "do the

9 OECD Recommendation of the Council on Artificial Intelligence, adopted 22.05.2019 (OECD Principles on AI); cf. OECD/LEGAL/0449, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; in German (unofficial translation) "Empfehlung des Rats zu künstlicher Intelligenz" available at: http://www.oecd.org/berlin/presse/Empfehlung-des-Rats-zu-kuenstlicher-Intelligenz.pdf.

10 *Stuart J. Russel*/*Peter Norving*, Artificial Intelligence – A Modern Approach, 3rd ed, 2016, 1. Other define the field of AI as "a field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – appear to be animals), and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – appear to be persons)." For this and a discus-sion of different approaches see *Selmer Bringsjord*/*Naveen Sundar Govindarajulu*, Artificial Intel-ligence, in *Edward N. Zalta* (ed.), Stanford Encyclopedia of Philosophy (SEP), Winter 2019 Ed.

11 *Stuart J. Russel*/*Peter Norving*, Artificial Intelligence – A Modern Approach, 3rd ed, 2016, 1.

12 *Stuart J. Russel*/*Peter Norving*, ibid., 2.

13 The famous and often quoted so-called Turing Test by *Alan M. Turing* is a behavioral intelligence test that shall provide an operational definition of intelligence. According to this test a programm passes the test if a human interrogator, after posing written questions via online typed messages for five minutes, cannot tell whether the written answers are given by a human being or a comput-er, cf. *Alan M. Turing*, Computing Machinery and Intelligence, Mind LIX, 1950, 433 et seq. (re-printed in *Margaret A. Boden* (ed.), The Philosophy of Artificial Intelligence, 1990, 40 et seq.); for a discussion see *Stuart J. Russel*/*Peter Norving*, Artificial Intelligence – A Modern Approach, 3rd ed, 2016, 1036 et seq.

14 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 477, 483.

15 An algorithm is a process (or program) that a computer can follow. It, for instance, defines a process to analyze a dataset and identify patterns in the data; in more general terms it can be desribed as a sequence of instructions that are carried out to transform the input to the output, see *John D. Kelleher*, Deep Learing, 2019, 7; *Ethem Alpaydin*, Machine Learning – The New AI, 2016, 16.

16 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 477.

right thing". Depending on the areas of use, AI key capabilities are natural language processing (speech recognition), reasoning, (learning, perception, and action (robotics). Especially learning[17] is a key ability of modern AI systems, [18] as for some problems it is unclear how to transform the input to the output.[19] This means that algorithms are developed that enable the machine to extract functions from a dataset to fulfill a certain task.[20] The socalled deep learning, the field of machine learning that focuses on deep neural networks,[21] is the central part of current AI systems if large datasets are available, as for face recognition on digital cameras[22] or in the field of medicine to diagnose certain illnesses.[23] Deep learning mechanisms that are able to improve themselves without human interaction and without rule-based programming already exist today.[24] As *John Kelleher* puts it: "Deep learning enables data-driven decisions by identifying and extracting patterns from large datasets".[25] It is not astonishing that since 2012 the number of new deep learning AI algorithms has grown exponential[26] but as the functional processes that generate the output are not clear (or at least hard to interpret) the problem of the complexity and opacity of algorithms that seem to be "black boxes" is obvious as well.[27]

2. Risks and Chances

The "black boxes" problem shows that it is important, if we think about AI regulation or governance, to look at the different risks and chances that can be linked to the development and use of AI systems. Questions of concern that are raised are related to our democratic order (news ranking algorithms, "algorithmic justice"), kinetics (autonomous cars and autonomous weapons), our economy and markets (algorithmic trading and pricing), and our society (conversational robots). A major and inherent risk if a system learns from data is that bias in AI systems can hardly be avoided. At least if AI learns from human-generated (text) data, they can or even will include health, gender or racial stereotypes.[28] Some claim, however, that there are better ways for reducing bias in AI than for reducing bias in humans, so AI systems may be or become less biased than humans.[29] Besides, there are risks of misuse, if AI systems are used to commit crimes, as for instance fraud.[30]

Another risk is that AI technologies have the potential for greater concentration of power. Those who are able to use this technology can become more powerful (corporations or governments),[31] and can influence large numbers of people (for instance to vote in a certain way). It was *Norbert Wiener* who wrote in 1954

> "(…) that such machines, though helpless by themselves, may be used by a human being or a block of human beings to increase their control over the rest of the race or that political leaders may attempt to control their populations by means not of machines themselves but through political techniques as narrow and indifferent to human possibility as if they had , in fact, been conceived mechanically."[32]

If we think about regulation, we must not forget the unintended and unanticipated negative and/or positive consequences of AI systems and that there might be a severe lack of predictability of these consequences.[33] The

17 The idea of a learning machine was discussed by *Alan M. Turing*, Computing Machinery and Intelligence, Mind LIX, 1950, 433 et seq. (reprinted in *Margaret A. Boden* (ed.), The Philosophy of Artificial Intelligence, 1990, 64 et seq.)

18 In general different types of feedback can be part of the machine learning process. There is unsupervised learning (no explicit feedback is given), reinforcement learning (the system learns based on rewards or "punishments"), and supervised learning, which means in order to teach a system what a tea cup is, you have to show it thousands of tea cups, cf. *Stuart J. Russel*/*Peter Norving*, Artificial Intelligence – A Modern Approach, 3rd ed, 2016, 706 et seq.

19 *Ethem Alpaydin*, Machine Learning – The New AI, 2016, 16 et seq.

20 *John D. Kelleher*, Deep Learing, 2019, 6; *Ethem Alpaydin*, Machine Learning – The New AI, 2016, 16 et seq.

21 *John D. Kelleher*, Deep Learing, 2019, 8.

22 *John D. Kelleher*, Deep Learing, 2019, 1: "Deep learning is the subfield of artificial intelligence that focuses on creating large neural network models that are capable of making accurate data-driven decisions." *Ethem Alpaydin*, Machine Learning – The New AI, 2016, 104: "With few as-sumptions and little manual interference, structures similar to the hierarchical cone are being au-tomatically learned from large amounts of data. (…) This is the idea behind deep neural networks where, starting from the raw input, each hidden layer combines the values in its preceding layer and learns more complicated functions of the input."

23 *Eric Topol*, Deep Medicine, 2019, 9 et seq., 16 et seq.

24 See *Yann LeCun* et al., Deep Learning, Nature 521 (2015), 436–444, available at: http://www.nature.com/nature/journal/v521/n7553/full/nature14539.html.

25 *John D. Kelleher*, Deep Learning, 2019, 4.

26 *Eric Topol*, Deep Medicine, 2019, 10.

27 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 478.

28 *Andrew Ng*, in *Martin Ford* (ed.), Architects of Intelligence, 2018, 20; Gutachten der Datenethikkommission, 2019, 167 f.

29 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 478.

30 *Stuart Russel*, Human Compatible – Artificial Intelligence and the Problem of Control, 253 et seq.

31 *W. Daniel Hills*, The First Machine Intelligences, in *John Brockman* (ed.), Possible Minds – 25 Ways of Looking at AI, 2019, 172, 173.

32 *Norbert Wiener*, The Human Use of Human Beings, 1954, 181.

33 *Stuart Russel*, Human Compatible – Artificial Intelligence and the Problem of Control, 103 et seq; *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 477 et seq.

use of AI will provide new and even better ways to improve our health system, to protect our environment and to allocate resources[34] However, plausible risk scenarios may show that the fear of the potential loss of human oversight is not per se irrational.[35] They support the call for a "human in the loop", that – for instance – a judge decides about the fate of a person, not an AI system, and a combatant decides about lethal or non-lethal force during an armed conflict, not an autonomous weapon. But to keep us as persons "in the loop" means that we need state based regulation stressing this as a necessary precondition at least in the areas where there are serious risks for the violation of human rights or human dignity. I agree with those who claim, that it is important to understand the properties of AI systems if we think about AI regulation and governance and that there is the need to look at the behavior of "black box" algorithms, similar to the behavior of animals, in real world settings.[36] My hypothesis is, that an AI system that serves human beings has to meet the "at least as good as a human being / human expert"[37] threshold. This sets even a higher threshold as the one that is part of the idea of "beneficial machines", defined as intelligent machines whose actions can be expected to achieve *our* objectives rather than *their* objectives.[38]

We also have to keep in mind the future development of AI systems and their interlinkage. I have spelled out so far features of so-called narrow AI or weak AI. Weak AI possesses specialized, domain specific, intelligence.[39] In contrast, Artificial General Intelligence (AGI) will possess general intelligence and strong AI could mean, as some claim, that AI systems "are actually thinking".[40] Whether there is a chance that AGI, and human-level or superhuman AI (the Singularity)[41] will be possible within our lifetime is uncertain.[42] It is not per se implausible to argue, as some scientists do, that intelligence explosion leads to a dynamically unstable system as smarter systems will have an easier time making themselves smarter[43] and that there will be a point beyond which it is impossible for us to make reliable predictions.[44] And it seems convincing that if superintelligent AI was possible it would be a significant risk for humanity.[45]

3. Current and Future AI Regulation

a. Bases

For regulative issues, the differentiation of narrow AI versus AGI might be helpful as a starting point. It is more convincing, however, to find categories that show the possible (negative) impact of AI systems to core human rights, human dignity and to constitutional rights, such as protection against discrimination, the right to life, the right to health, the right to privacy, and the right to take part in elections, etc.[46] From this perspective, even developments such as a fast take-off scenario, which means that an AGI system becomes super-intelligent because of

34 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 478.

35 Stressing the need to analyze risks, cf. *Max Tegmark*, Let's Aspire to More Than Making Our-selves Obsolete, in *John Brockman* (ed.), Possible Minds – 25 Ways of Looking at AI, 2019, 76 et seq.; *Stuart Russel*, Human Compatible – Artificial Intelligence and the Problem of Control, 103 et seq.

36 *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 478.

37 Depending on the area the AI system is deployed the system has to be measured against the human expert that usually is allowed to fulfil a task (as for instance an AI diagnosis system). This differs from the view of the German Datenethikkommission as the commission argues that there is an ethical obligation to use AI systems if they fulfil a certain task better as an human, cf. Gutachten der Datenethikkommission, 2019, 172.

38 *Stuart J. Russel*, Human Compatible, 2019, pp. 172 et seq.

39 Some claim that weak AI means that AI driven machines act "as if they were intelligent", cf. *Stuart J. Russel*/*Peter Norvig*, Artificial Intelligence – A Modern Approach, 3rd ed, 2016, 1035.

40 *Stuart J. Russel*/*Peter Norvig*, ibid., 1035; *Murray Shanahan*, The Technological Singularity, 2015, 3.

41 The term "the Singularity" was coined in 1993 by the computer scientist and author *Vernon Vinge*; he was convinced that "[w]ithin thirty years, we will have the technological means to create superhuman intelligence," and he concluded: "I think it's fair to call this event a singularity ("the Singularity" for the purpose of this paper)." See Vernor Vinge, The Coming Technological Singularity: How to Survive in the Post-Human Era, in *Geoffrey A. Landis* (ed.), Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace (1993), 11, 12 (NASA Publication CP-10129), available at: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf.

42 *Stuart J. Russel*, The Purpose Put into the Machine, in *John Brockman* (ed.), Possible Minds: 25 Ways of Looking at AI, 2019, 20 et seq., 26. Some experts predict that superhuman intelligence will happen by 2050, see e.g., *Ray Kurzweil*, The Singularity is Near, 2005, 127; for more forecasts, see *Nick Bostrom*, Superintelligence, Paths, Dangers, Strategies, 2014, at 19–21. 43 *Eliezer Yudkowsky*, Artificial Intelligence as a positive and negative factor in global risk, in *Nick Bostrom*/*Milan Ćirković* (eds.), Global Catastrophic Risks, 2011, at 341.

43 *Eliezer Yudkowsky*, Artificial Intelligence as a positive and negative factor in global risk, in Nick Bostrom/Milan Ćirković (eds.), Global Catastrophic Risks, 2011, at 341.

44 *Max Tegmark*, Will There Be a Singularity within Our Lifetime?, in *John Brockman* (ed.), What Should We Be Worried About?, 2014, 30, 32.

45 *Stuart J. Russel*, The Purpose Put into the Machine, in *John Brockman* (ed.), Possible Minds: 25 Ways of Looking at AI, 2019, 26.

46 For a similar approach, however less based in the risks for the violation of human rights, see Gutachten der Datenethikkommission, 2019, 173.

a recursive self-improvement cycle,[47] that are difficult to predict, must not be neglected as we can think about how to frame low probability high impact scenarios in a proportional way.[48]

b. Sector-Specific Rules and Multilevel Regulation

When speaking about governance and regulation, it is important to differentiate between rules that are legally binding on the one hand and non-binding soft law, on the other hand. In the area of international, European Union, and national law, we see that at least parts of AI-driven technology are covered by existing sector-specific rules.

(1) AI Systems Driven by (Big) Data

The General Data Protection Regulation (GDPR)[49] aims to protect personal data[50] of natural persons (art. 1 (1) GDPR) and applies to the processing of this data even by wholly automated means (art. 2 (1) GDPR).[51] The GDPR requires an informed consent[52] of the consumer if somebody wants to use his or her data. It can be seen as sector-specific law governing AI systems as AI systems often make use of large amounts of personal data. The general principles that are laid down for – inter alia – the processing of personal data (including lawfulness, fairness and transparency[53]) and the collection of personal data (purpose limitation) in art. 5 GDPR are applicable with regard to AI systems,[54] and have to be implemented via appropriate technical and organizational measures by the controller (art. 25 GDPR).[55] According to art. 22 GDPR we, as data subjects, have the right "not to be subject to a decision based solely on automated processing" that produces legal effects concerning the data subject or similarly affects him or her.[56] Substantive legitimacy of this regulations is given because the GDPR is in coherence with the human rights that bind EU organs and can be reviewed and implemented by the European Court of Justice and the German Constitutional Court,[57] especially art. 8 of the Charter of Fundamental Rights of the European Union (EUChHR)[58] that lays down the protection of personal data.[59] Like every regulation and law, the GDPR has lacunae, and there might be relevant lacunae in the area of AI-driven technology, as for instance, with regard to brain data that is used for consumer technology.[60] The decisive question is whether all relevant aspects of brain data protection are already

47 *Andrew Ng*, in *Martin Ford* (ed.), Architects of Intelligence, 2018, 202.

48 For a governance framework of superintelligent AI as an existential risk, see *Silja Voeneky*, Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks, in *Silja Voeneky*/*Gerald Neuman* (eds.), Human Rights, Democracy, and Legitimacy in Times of Disorder, 2018, 160 et seq. 49 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27.04.2016 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, in force since 25.05.2018, cf. OJEU L119/1, 04.05.2016.

50 Art. 4 (1) GDPR: "'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".

51 However, art. 2 (2) lit. c and d GDPR excludes from the material scope the processing as defined in art. 4 (2) GDPR of personal data by a natural person in the course „of a purely personal or household activity", and by the competent authorities for the purposes inter alia „of the prevention (…) or prosecution of criminal offences".

52 Cf. art. 7, art. 4 (11) GDPR: " (…) 'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relat-ing to him or her;".

53 See as well art. 12 GDPR.

54 See art. 6 (4) GDPR.

55 With regard to the responsible and accountable person or entity ("the controller" according to art. 4 (7) GDPR) and further duties of the controller see art. 5 (2) ("accountability"), art. 32 ("security of processing") and art. 35 GDPR ("data protection impact assessment"). For a discussion in Germany how to apply the GDPR to AI systems see, inter alia, the Entschließung der 97. Konferenz der unabhängigen Datenschutzaufsichtbehörden des Bundes und der Länder, 03.04.2019 ("Hambacher Erklärung zur Künstlichen Intelligenz"), available at: https://www.datenschutzkonferenz-online.de/media/en/20190405_hambacher_erklaerung.pdf. For the claim that there is a need for a new EU Regulation for AI systems, see the Gutachten der Datenethikkommission, 2019, 180 proposing a "EU-Verordnung für Algorithmische Systeme" (EUVAS).

56 See as well Gutachten der Datenethikkommission, 2019, 191 et seq.

57 The German Constitutional Court declared to be competent to review the application of national legislation on the basis of the rights of the Charter of Fundamental Rights of the European Union even in an area that is fully harmonized according to EU law, cf. BVerfG, Decision of 06.11.2019, 1 BvR 276/17, Right to be forgotten II.

58 Cf. OJEC C 364/1, 18.12.2000.

59 Art. 8 EUChHR: "Protection of personal data
(1) Everyone has the right to the protection of personal data concerning him or her.
(2) Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
(3) Compliance with these rules shall be subject to control by an independent authority."

60 Cf. *Oscar Schwartz*, Mind-Reading Tech? How Private Companies could gain Access to Our Brains, The Guardian, 24.10.2019, online available at: https://www.theguardian.com/technology/2019/oct/24/mind-reading-tech-private-companies-access-brains.

covered by the protection of health data (art. 4 (15) GDPR) or biometric data (art. 4 (14) GDPR) that are defined in the regulation.[61]

(2) AI Systems as Medical Devices

Besides, there is EU Regulation on Medical Devices (MDR),[62] which governs certain AI-driven apps in the health sector and other AI-driven medical devices, for instance, in the area of neurotechnology.[63] And again, one has to ask whether this regulation is sufficient to protect the human dignity, life and health of consumers, as the impact on human dignity, life and health might be more fareaching than the usual products that were envisaged by the drafters of the regulation. Although the new EU medical device regulation was adopted in 2017, it includes a so-called scrutiny process[64] for high-risk products (certain class III devices), which is a consultation procedure prior to market approval. It is not a preventive permit procedure, differing from the permit procedure necessary for the market approval of new medicine (medicinal products), as there is a detailed regulation at the national and even more at the European Union level,[65] including a new Clinical Trial Regulation.[66] That the preventive procedures differ whether the object of the relevant laws is a "medical device" or a "medicinal product" is not convincing, if the risks involved for human health for a consumer are the same when comparing new drugs and certain new medical devices, as – for instance – new neurotechnology.

(3) AI Systems as (Semi-)Autonomous Cars

Sector-specific (top-down) regulation is already in force when it comes to the use of (semi-)autonomous cars. In Germany, the relevant national law was amended in 2017,[67] before the competent federal ethic commission published its report,[68] in order to include new highly or fully automated systems (§ 1a, § 1b and § 63 StVG). § 1a (1) StVG states that the operation of a car by means of a highly or fully automated driving function is permissible, provided the function is used for its *intended purpose*.[69] However, what "intended purpose" means must be defined by the automotive company. Therefore § 1a (1) StVG means a dynamic reference to the private standard-setting by a corporation that seems to be rather vague[70] especially if you think about the rule of law and the principle of "Rechtsklarheit", which means that legal rules have to be clear and understandable.[71] It is even true with regard to the applicable international treaties that sector-specific law can be amended and changed (even at the international level) if it is necessary to adapt the old rules to now AI-driven systems. The UN/ECE 1958 Agreement[72] was amended in 2017 and 2018 (the

61 To discuss this in detail is beyond the scope of this paper but it is one area of research of the Sal-tus-FRIAS Responsible AI Research Group the author is part of.

62 Regulation (EU) 2017/745 of the European Parliament and of the Council of 05.04.2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, OJEU L117/1, 05.05.2017. It came into force May 2017, but medical devices will have a transition time of three years (until May 2020) to meet the new requirements.

63 Art. 2 MDR. „ (…) 'medical device' means any instrument, apparatus, appliance, software, im-plant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes: (…)". For exemptions see, however, art. 1 (6) MDR.

64 Cf. art. 54, 55, art. 106 (3), Annex IX Section 5.1, Annex X Section 6 MDR.

65 Cf. the Pharmaceutical legislation for medicinal products of human use, Vol. 1, including different Directives and Regulations, available at: https://ec.europa.eu/health/documents/eudralex/vol-1_de.

66 §§ 21 et seq. Gesetz über den Verkehr mit Arzneimitteln (Arzneimittelgesetz, AMG), BGBl. I, 1626; Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16.04.2014 on clinical trials on medicinal products for human use, OJEU L 158/1, 27.05.2014.

67 Art. 1 Achtes Gesetz zur Änderung des Straßenverkehrsgesetzes (8. StVGÄndG), 16.06.2017, BGBl. I 1648.

68 Ethik-Kommission „Automatisiertes und vernetztes Fahren" des Bundesministeriums für Verkehr und digitale Infrastruktur, Report June 2017, available at: https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile.

69 § 1a (1) StVG: „Der Betrieb eines Kraftfahrzeugs mittels hoch- und vollautomatisierter Fahrfunktion ist zulässig, wenn die Funktion bestimmungsgemäß verwendet wird."

70 This seems true even if the description of the intended purpose and the level of automation shall be „unambigous" according to rationale of the law maker, cf. BT-Drucks., 18/11300, 20: "Die Systembeschreibung des Fahrzeugs muss über die Art der Ausstattung mit automatisierter Fahrfunktion und über den Grad der Automatisierung unmissverständlich Auskunft geben, um den Fahrer über den Rahmen der bestimmungsgemäßen Verwendung zu informieren."

71 *Bernd Grzeszick*, art. 20, in *Roman Herzog*/*Rupert Scholz*/*Matthias Herdegen*/*Hans Klein* (eds.), *Maunz*/*Dürig* Grundgesetz-Kommentar, para. 51–57.

72 Agreement Concerning the Adoption of Harmonized Technical United Nations Regulations for Wheeled Vehicles, Equipment and Parts which can be Fitted and/or be Used on Wheeled Vehicles and the Conditions for Reciprocal Recognition of Approvals Granted on the Basis of these United Nations Regulations.

UN Regulations No. 79[73] and No. 13-H[74]) to have a legal basis for the use of (semi-)autonomous cars.[75]

The examples mentioned above show that detailed, legally binding laws and regulations are already in force to regulate AI systems at the international, European, and national level. According to this, the "narrative" is not correct which includes the claim that (top-down) state-based regulation lags (or: must lag) behind the technical development, especially in the area of a fast-moving disruptive technology as AI. It seems rather convincing to argue instead that whether there is meaningful regulation in the field of AI depends on the political will to regulate AI systems at the national, European, and international level.

(4) AI Systems as (Semi-)Autonomous Weapons

The political will to regulate will depend on the interest(s) and preferences of states, especially with regard to economic goals and security issues as in most societies (democratic or undemocratic) there seems broad consensus that economic growth of the national economy *is* a (primary) aim and providing national security is the most important legitimate goal of a state. This might explain why there are at the international level – at least until now – areas where there is no consensus to regulate AI systems as a regulation is seen as a limiting force for economic growth and/or national security.

This is obvious with regard to (semi-)autonomous weapons. Though a Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS) was established in 2016 and has met in Geneva since 2017 convened through the Conference on Certain Conventional Weapons (CCW) and a report of the 2019 session of the GGE is published[76] there are only guiding principles affirmed by the Group.[77] These guiding principles stress, inter alia, the need for accountability (lit. b and d),[78] and risk assessment measures as part of the design (lit. g). However, there is no agreement on a meaningful international treaty, and it is still disputed whether the discussion within the GGE should be limited to fully autonomous systems.[79]

The mostly state-driven discussions at the CCW have shown that some States are arguing for a prohibition as

73 Addendum 78: UN Regulation No. 79 Rev. 3, ECE/TRANS/WP.29/2016/57 ECE/TRANS/WP.29/2017/10 (as amended by paragraph 70 of the report ECE/TRANS/WP.29/1129), 30.11.2017, "Uniform provisions concerning the approval of vehicles with regard to steering equipment":
"2.3.4.1. 'Automatically commanded steering function (ACSF)' means a function within an electronic control system where actuation of the steering system can result from automatic evaluation of signals initiated on-board the vehicle, possibly in conjunction with passive infrastructure features, to generate control action in order to assist the driver.
2.3.4.1.1. 'ACSF of Category A' means a function that operates at a speed no greater than 10 km/h to assist the driver, on demand, in low speed or parking manoeuvring.
2.3.4.1.2. 'ACSF of Category B1' means a function which assists the driver in keeping the vehicle within the chosen lane, by influencing the lateral movement of the vehicle.
2.3.4.1.3. 'ACSF of Category B2' means a function which is initiated/activated by the driver and which keeps the vehicle within its lane by influencing the lateral movement of the vehicle for extended periods without further driver command/confirmation.
2.3.4.1.4. 'ACSF of Category C' means, a function which is initiated/activated by the driver and which can perform a single lateral manoeuvre (e.g. lane change) when commanded by the driver.
2.3.4.1.5. 'ACSF of Category D' means a function which is initiated/activated by the driver and which can indicate the possibility of a single lateral manoeuvre (e.g. lane change) but performs that function only following a confirmation by the driver.
2.3.4.1.6. 'ACSF of Category E' means a function which is initiated/activated by the driver and which can continuously determine the possibility of a manoeuvre (e.g. lane change) and complete these manoeuvres for extended periods without further driver command/confirmation."

74 Addendum 12-H: UN Regulation No. 13-H, ECE/TRANS/WP.29/2014/46/Rev.1 and ECE/TRANS/WP.29/2016/50, 05.06.2018, "Uniform provisions concerning the approval of passenger cars with regard to braking": 2.20. "'Automatically commanded braking' means a function within a complex electronic control system where actuation of the braking system(s) or brakes of certain axles is made for the purpose of generating vehicle retardation with or without a direct action of the driver, resulting from the automatic evaluation of on-board initiated information."

75 To understand the relevance of these regulations in a multi-level regulation system one has to take into account that other international, European national provisions refer directly or indirectly to the UN/ECE Regulations, cf. e.g. art. 8 (5bis) and art. 39 of the Vienna Convention on Road Traf-fic; art. 21 (1), 29 (3), 35 (2) of the European Directive 2007/46/EC ("Framework Directive"); § 1a (3) StVG.

76 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (CCW), Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 25.–29.03.2019 and 20.–21.08.2019, Report of the 2019 session, CCW/GGW.1/2019/3, 25.09.2019, available at: https://undocs.org/en/CCW/GGE.1/2019/3.

77 Ibid., Annex IV, 13 et seq.

78 Ibid., Annex IV: (b) "Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system; (…)
(d) Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, includ-ing through the operation of such systems within a responsible chain of human command and control;".

79 For this view and a definition see working paper (WP) submitted by the Russian Federation, CCW/GGE.1/2019/WP.1, 15.03.2019, para. 5: "unmanned technical means other than ordnance that are intended for carrying out combat and support missions without any involvement of the operator", expressly excluding unmanned aerial vehicles as highly automated systems.

part of a new international treaty, like Austria, yet other States, like Russia[80] and the US,[81] are stressing the advantages[82] of the development and use of (semi-)autonomous weapons. Germany and France[83] do not support an international treaty but opted for a soft law code of conduct with regard to framing the use of those weapons.[84]

Besides, key elements of a governance regime of (semi-)autonomous weapons are unclear. What is meant by "human control over the operation of such systems" is discussed even if it is stated that this is an important limiting factor by a state. Russia, for instance, argues that

"the control system of LAWS should provide for intervention by a human operator or the upper-level control system *to change the mode of operation of such systems, including partial or complete deactivation".*[85]

With this, Russia eliminates meaningful human control as a necessary precondition to use (semi-)autonomous weapons. The "human in the loop" as a last resort of using lethal weapons and the subject of responsibility – with the last resort to convict somebody as a war criminal – is replaced by the *upper-level control system* that might be another AI system.

(5) First Conclusion

The examples mentioned above show the loopholes of the international regulation of AI systems, although there are specific rules in place in some areas, mostly at the European level. But more importantly that there is no coherent, general, or universal international regulation of AI as part of the international hard law. Although there are lacunae in other areas as well (thus far no international treaty on existential and global catastrophic risks and scientific research exists) this widespread international non-regulation of AI research and development is different from other fields of fast moving technological progress: biotechnology. In the field of biotechnology there are a treaties, like the Biological Weapons Convention (BWC),[86] the Convention on Biological Diversity, the Cartagena Protocol on Biosafety,[87] and the Kuala Lumpur Liability Protocol[88] that are applicable in order to prohibit research that is not aimed at peaceful purposes or to dimisih risks related to the genetic modification of living organisms. Therefore, it is important to look closer to the first attempt to adopt general AI principles at the international level as part of the international soft law.[89]

II. OECD AI Recommendations as International Soft Law

1. Basis and Content

The OECD issued recommendations on AI in 2019[90] and 43 States have adopted these principles[91] including relevant actors in the field of AI as the US, South Korea, Japan, UK, France, and Germany, and States that are not members of the OECD. The recommendations were drafted with the help of an expert group (AIGO) that

80 WP submitted by the Russian Federation, CCW/GGE.1/2019/WP.1; para. 2: "The Russian Federation presumes that potential LAWS can be more efficient than a human operator in addressing the tasks my minimizing the error rate. (…)."

81 WP submitted by the USA, CCW/GGE.1/2019/WP.5, 28.03.2019, para. 2 lit. c: "Emerging technologies in the area of LAWS could strengthen the implementation of IHL, by, inter alia, reducing the risk of civilian casualties, facilitating the investigation or reporting of incidents involving potential violations, enhancing the ability to implement corrective actions, and automatically generating information on unexploded ordnance."; cf. as well ibid., para. 15.

82 WP submitted by the Russian Federation, CCW/GGE.1/2019/WP.1, para. 10: "The Russian Federation is convinced that the issue of LAWS is extremely sensitive. While discussing it, the GGE should not ignore potential benefits of such systems in the context of ensuring States' national security. (…)".

83 WP submitted by France, CCW/GGe.2/2018/WP.3, stressing inter alia the principles of command responsibility, ibid. para. 6, stressing a "central role for human command in the use of force" (para. 12): "(…) In this regard, the command must retain the ability to take final decisions regarding the use of lethal force including within the framework of using systems with levels of autonomy or with various artificial intelligence components."

84 Even the German Datenethikkommission stresses that there is not per se a "red line" with regard to autonomous weapons as long as the killing of human beings it not determined by an AI system, Gutachten der Datenethikkommission, 2019, 180.

85 WP submitted by the Russian Federation, CCW/GGE.1/2019/WP.1, para. 7. For a different approach see the ICRC Working Paper on Autonomy, AI and Robotics: Technical Aspects of Hu-man Control, CCW/GGE.1/2019/WP.7, 20.08.2019.

86 16.12.1971, 1015 U.N.T.S. 163, entered into force 26.03.1975. The BWC allows research on biological agents for preventive, protective or other peaceful purposes; however this treaty does not provide sufficient protection against the risks of misuse of research because research conducted for peaceful purposes is neither limited nor prohibited.

87 29.01.2000, 2226 U.N.T.S. 208, entered into force 11.09.2003.

88 The Nagoya-Kuala Lumpur Supplementary Protocol on Liability and Redress to the Cartagena Protocol on Biosafety, 15.10.2010, entered into force 05.03.2018.

89 The term "international soft law" is understood in this paper to cover rules that cannot be attributed to a formal legal source of public international law and that are, hence, not directly legally binding but have been agreed upon by subjects of international law (i.e. States, international organizations) that could, in principle, establish international hard law; for a similar definition see Daniel Thürer, Soft Law, in Rüdiger Wolfrum (ed.), Max Planck Encyclopedia of Public International Law, 2012, Vol. 9, 271, para. 8. The notion does not include private rule making by corporations (including codes of conduct) or mere recommendations by stakeholders, non-governmental organisations and other private entities.

90 See above note 9.

91 Cf. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

consists of 50 members from – as the OECD writes – governments,[92] academia, business, civil society etc., including IBM, Microsoft, Google, Facebook, DeepMind, as well as invited experts from MIT.[93] The OECD claims that these Principles will be a global reference point for trustworthy AI.[94] It refers to the notion of trustworthy AI, as did the High-level Expert Group on AI (AI HLEG) set up by the EU, which published Ethics Guidelines on AI in April 2019 listing seven key requirements that AI systems shall meet to be trustworthy.[95]

The OECD recommendations state and spell out five complementary value-based "principles for responsible stewardship of trustworthy AI" (section1):[96] these are inclusive growth, sustainable development and well-being (1.1); human-centered values and fairness (1.2.); transparency and explainability (1.3.); robustness, security and safety (1.4.); and accountability (1.5.). In addition, AI actors – meaning those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI[97] – should respect the rule for human rights and democratic values (1.2. lit. a). These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights. But the wording of the principles is very soft. For instance, AI actors should implement
"mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of the art" (1.2. lit. b).

The recommendation about transparency and explainability (1.3.) has only slightly more substance. It states that AI actors

"[…] should provide meaningful information, appropriate to the context, and consistent with the state of art […] *(iv.) to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, rec-ommendation or decision.*"

Additionally, it states that

"AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach *to each phase of the AI system lifecycle, on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias." (1.4 lit. c).*

If we think that discrimination and unjustified biases are one of the key problems of AI,[98] asking for a risk management approach to avoid these problems does not seem to be sufficient as a standard of AI actor (corporation) due diligence.

And the wording with regard to accountability is soft as well (1.5):

"AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context and consistent with the state for the art."
This does not mean and does not mention any legal liability or legal responsibility."

2. (Dis-)Advantages and Legitimacy

The OECD recommendations show some of the advantages and disadvantages that we see in the area of international soft law. The *advantages* are that they can be drafted in a short period of time (the working group started in 2018); that they can include experts from the relevant fields and state officials; that they can spell out and identify an existing overlapping consensus of member states, here the OECD member states; and that they might develop some kind of normative force even if they are not legally binding as an international treaty.[99]

However, the *disadvantages* of the OECD recommendations are obvious as well. *Firstly*, the basis for the *procedural legitimacy* is unclear as to which experts are allowed to participate is not entirely clear. In the field of AI, experts are employed, paid, or closely linked to AI corporations[100] hence, the advice they give is not (entirely) independent. If an International Organisation (IO) or State one wants to enhance procedural legitimacy for AI recommendations, one should rely on different groups: one of the independent experts with no (financial) links to corporations, one of the experts working for corporations, and a third group consisting of civil society and

92 Germany did send one member (Policy Law: Digital Work and Society, Federal Ministry for La-bour and Social Affairs), Japan two, as well as France, and the European Commission; South Korea did send three members, as the USA (US Department of State, US Department of Commerce; US National science Foundation).

93 Cf. https://www.oecd.org/going-digital/ai/oecd-aigo-membership-list.pdf.

94 Cf. OECD Website: What are the OECD Principles on AI?, https://www.oecd.org/going-digital/ai/principles/.

95 These are: 1. Human agency and oversight; 2. Technical robustness and safety 3. Privacy and data governance; 4. Transparency; 5. Diversity, non-discrimination and fairness; 6. Societal and environmental well-being 7. Accountability.

96 An AI system is defined as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy." Cf. I OECD AI Recommendations.

97 Ibid., I OECD AI Recommendations.

98 See above at note 28. See as well Gutachten der Datenethikkommission, 2019, 194.

99 *Silja Vöneky*, Recht, Moral und Ethik, 2010, 284 et seq.

100 See above at note 93.

NGO members. States or IO could then compare the recommendations, discuss the differences, and choose or combine the one most convincing.

Secondly, we have to discuss the *substantive legitimacy* because the OECD recommendations do not stress the responsibility of governments to protect human rights in the area of AI. They include only five recommendations to policymakers ("adherents", section 2) that shall be implemented in national policies and international cooperation consistent with the principles mentioned above. These include investing in AI research and development (2.1), fostering a digital ecosystem for AI (2.2), shaping an enabling policy environment for AI (2.3), building human capacity and preparing for labor market transformation (2.4), and international cooperation for trustworthy AI (2.5).

3. Second Conclusion

As a conclusion of this second part one could state that the OECD recommendations lower the threshold too far and shift the focus too far away from States as main actors of the international community and as those obliged to protect human rights[101] towards private actors. This is a major disadvantage because although these recommendations exist, it is still unclear what state obligations can be deduced from legally binding human rights – including the relevant human rights treaties and rules of customary law – with regard to the governance of AI. Besides, the recommendations that address private actors and their responsibilities are drafted in a language that is too soft and vague. As a result, I argue that the OECD Recommendations could and should have been more meaningful with regard to standards of due diligence and responsibility in the age of AI for private actors and – even more – with regard to state duties to protect human rights. The latter aspect might even lead to a trend to undermine state duties to protect human rights in times of AI – and this could undermine the relevance of human rights regarding AI regulation as a whole.

III. Legitimacy, Human Rights and AI Regulation

The question to be answered in this third part is: why are human rights decisive with regard to the regulation of AI, and how can we defend the link between legitimacy and human rights in the field of AI regulation?

1. Legitimacy

I start with the notion of legitimacy. As I have written before, legitimacy should be viewed primarily as a normative, not a descriptive, concept: [102] It refers to standards of justification of governance, regulation and obligations. Hence, legitimate governance or regulation means that the guiding norms and standards have to be justifiable in a supra-legal way (i.e. they possess *rational acceptability*). If we think about international regulation, it seems fruitful to link the notion of "legitimate regulation" to the existing legal order of public international law. Without saying that legality is sufficient for legitimacy, I argue that guiding norms and standards have to be coherent with existing international law insofar as the international law reflects moral (i.e. justified) values.[103]

2. Ethical Paradigms

We have to state that there are different ethical paradigms that can justify regulation in the field of AI in a supra legal way. One is the human rights-based approach that can be considered a deontological concept,[104] as the rightness or wrongness of conduct is derived from the character of the behavior itself.[105] Another approach is utilitarianism, which can be described as the doctrine which states that "one should perform that act, among those that on the evidence are available to one, that will most probably maximise benefits".[106] It seems important to note that the different normative ethical theories are

101 See below Part III.

102 The arguments at part III. 1.-3. were published in my paper Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks, in *Silja Voeneky/Gerald Neuman* (eds.), Human Rights, Democracy, and Legitimacy in Times of Disorder, 2018, 149.

103 For the basis on the concept and notion of "legitimacy", see *Silja Voeneky*, Recht, Moral und Ethik, 2010, 130–162. For discussion of the legitimacy of international law, see *Allen Buchanan*, The Legitimacy of International Law, in *Samantha Besson/John Tasioulas* (eds.), The Philosophy of International Law, 2010, 79–96; *John Tasioulas*, Legitimacy of International Law, in *Samantha Besson/John Tasioulas* (eds.), The Philosophy of International Law, 2010, at 97–116.

104 A deontological theory of ethics is one which holds that at least some acts are morally obligatory regardless of their consequences, see *Robert G. Olson*, in *Paul Edward* (ed.), The Encyclopedia of Philosophy, 1967, 1–2, 343.

105 Which means that these views maintain that "it is sometimes wrong to do what produces the best available outcome overall" as these views incorporate "agent-centred restrictions," see *Samuel Scheffler*, The Rejection of Consequentialism, 1994, 2.

106 On "direct" and "act" utilitarianism, see *Richard B. Brandt*, Facts, Values, and Morality, 1996, 142; for the notion of act-consequentialism and classical utilitarianism see *Samuel Scheffler*, supra note 105, at 2-3; for an overview see *John C. Smart*, Utilitarianism, in *Paul Edward* (ed.), The Encyclopedia of Philosophy 1967, 7–8, 206.

based on reasonable grounds (i.e. they possess rational acceptability),[107] and one cannot decide whether there is a theory that clearly trumps the others. Therefore, in looking for standards that are the bases of legitimate regulation of AI systems, it is not fruitful to decide whether one normative ethical theory is in general terms the most convincing one, but rather which ethical paradigm seems to be the most convincing in regard to the specific questions that we have to deal with when framing AI systems.

3. Human Rights-based AI Regulation

As I have argued before with regard to the regulation of existential risks, [108] I argue that AI regulation and governance should be based on human rights, more precisely on legally binding human rights. Other ethical approaches shall not be ruled out as far as they are compatible with human rights. But I reject views that argue that utilitarian standards should be the primary standard to measure the legitimacy of an AI regulative regime.[109] The arguments supporting this claim are the following:
To regulate AI is a global challenge. Hence, it would be a major deficit not to rely on human rights. They are part of existing international law. They are not only rooted in the moral discourse as universal values, but they also bind many, or even all States (as treaty law or customary law), and they can be implemented by courts or other institutional means, laid down in human right treaties, such as the European Convention on Human Rights (ECHR)[110] and the International Covenant on Civil and Political Rights (ICCPR). The latter is a universal human rights treaty that is binding on more than 170 States Parties,[111] including major AI relevant actors, like the USA.

What seems to be even more important is that when we turn to a human rights framework, we see that international legal human rights make it possible to spell out the decisive values that must be taken into account for assessing different AI-research, -development and -deployment scenarios. In the area of AI research freedom of research is decisive as a legally binding human right, entailed in the rights of freedom of thought and freedom of expression that are laid down in the CCPR as an international universal human rights treaty. However, this freedom is not absolute: The protection – for instance – of life and health of human beings, of privacy and against discrimination are legitimate aims that can justify *proportional limitations* of this right.[112] The human rights framework, therefore, stresses that there exists a need to find proportional limitations in the field of AI research if there are dangers or risks[113] for human life and health or

107 In order to argue this way we have to answer the question what our criteria of rational acceptability are. My answer is based on the arguments by the philosopher *Hilary Putnam* that our criteria of rational acceptability are, inter alia, coherence, consistency, and relevance; that "fact (or truth) and rationality are interdependent notions" but that, nevertheless, no neutral understanding of rationality exists as the criteria of "rational acceptability rest on and presuppose our values", and the "theory of truth presupposes theory of rationality which in turn presupposes our theory of good". Putnam concluded that the theory of the good is "itself dependent upon assumptions about human nature, about society, about the universe (including theological and metaphysical assumptions)." See *Hilary Putnam*, Reason, Truth and History, 1981, 198, 201, 215.

108 This and the arguments at III.2. and 3. were published in my paper Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks, in *Silja Voeneky*/*Gerald Neuman* (eds.), Human Rights, Democracy, and Legitimacy in Times of Disorder, 2018, 151 et seq.

109 In many cases, neither the risks nor the benefits of AI research and development can be quantified; the risk of misuse of AI systems by criminals, mentioned above, cannot be quantified; the unclear or unpredictable benefits of basic AI research cannot be quantified either – nevertheless, basic research may often be the necessary condition in order to achieve benefits for human beings in the long run. These are drawbacks of a utilitarian risk-benefit approach for some of the AI scenarios described above. For the lack of predictability surrounding the consequences of AI, cf. *Iyad Rahwan*/*Manuel Cebrian*/*Nick Obradovich* et al., Machine behaviour, Nature 568 (2019), 477. For a general discussion of the human rights approach versus utilitarianism see *Herbert L. A. Hart*, Between Utility and Rights, Colum. L. Rev. 79 (1979), 828. For a discussion of a combination of utilitarianism and other value based approaches (autonomy, diversity) and reference to the Universal Declaration of Human Rights for the codification of moral principles applicable to future AI, see *Max Tegmark*, Life 3.0, 2017, 271–75.

110 The International Covenant on Civil and Political Rights adopted by G.A. Res. 2200A (XXI), 16.12.1966, entered into force 23.03.1976, 999 U.N.T.S. 171, and the European Convention on Human Rights, adopted by the Members of the Council of Europe, 04.11.1950, available at: http://www.echr.coe.int/Documents/Convention_ENG.pdf.

111 Art. 18 ICCPR, 19; art. 9, 10 ECHR. A different approach is taken, however, in the Charter of Fundamental Rights of the European Union, art. 13 (Freedom of the arts and sciences). There it is expressly laid down that "The arts and scientific research shall be free of constraint. Academic freedom shall be respected." Similar norms are included in national constitutions, see e.g. Grundgesetz für die Bundesrepublik Deutschland, art. 5 (3) (23.05.1949) which states that "Arts and sciences, research and teaching shall be free. The freedom of teaching shall not release any person from allegiance to the constitution."

112 The legitimate aims for which the right of freedom of expression and the right of freedom of science can be limited according to the International Covenant on Civil and Political Rights and the European Convention on Human Rights are even broader. See art. 19 (3) ICCPR, art. 10 (2) ECHR.

113 Risk can be defined as a risk is an "unwanted event which may or may not occur", see *Sven O. Hansson*, Risk, in *Edward N. Zalta* (ed.), Stanford Encyclopedia of Philosophy, available at: https://plato.stanford.edu/entries/risk/. There is no accepted definition of the term in public international law; it is unclear how—and whether—a "risk" is different from a "threat," a "danger" and a "hazard," see *Grant Wilson*, Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law, Virginia Environmental L.J. 31 (2013), 307, 310.

privacy. What limits to the freedom of research are justified depends on the probability of the realization of a risk[114] and the severity of the possible harm.

Therefore, demands of rational risk-benefit assessment can and should be part of the interpretation of human rights, as there is the need to avoid disproportionate means in order to minimize risks even in low/unknown probability cases: What proportionality means is linked to the risks and benefits one can reasonably anticipate in the area of AI. To do a risk-benefit assessment of the AI system in question, as far as this is possible and rational, therefore is an important element in implementing the human rights framework.

Besides, even so-called first generation human rights, as laid down in the CCPR, oblige States not only to respect, but also to protect the fundamental rights of individuals.[115] They state that States parties are obliged by international human rights treaties to take appropriate (legal) measures to protect inter alia the life and health of individuals.[116] And although there is wide discretion for States to protect human rights, measures must not be ineffective.

Last but not least, a human rights-based approach requires procedural rights for individuals to participate in the making of decisions that affect them in the area of AI-developments. To rely on human rights mean that we have to spell out in more detail, how to enhance procedural legitimacy.

These arguments might show that the core of the regulation and governance problem – that AI systems should serve us as human beings and not the other way around – can be expressed best on the basis of a human rights framework. It is correct that human rights law, even the right to life, is not aiming to protect humanity, but aiming to protect individuals.[117] However, humanity consists of us as individuals. Even if we are not arguing that human rights protect future generations, we may not neglect that individuals born today can have a life expectancy of more than 70 years in many States, and these individuals are protected by human rights law. Hence, it seems consistent with the object and purpose of human rights treaties that we view human rights law, and the duty of States towards human beings because of human rights, in a 70 year period.

IV. Future AI Regulation

In this paper, I spell out what the deficiencies of current AI regulations (including international soft law) are (part I and II), and I argue why international law, and international human rights are and should be the basis for a legitimate global AI regulation and risk reduction regime (part III). This approach makes it possible to develop rules with regard to AI systems in coherence with relevant and morally justified values of a humane world order that is aiming for future scientific and technological advances in a responsible manner, including the human right to life, the right to non-discrimination, the right to privacy and the right to freedom of science.

However, this is only a first step as current human rights norms and treaties are a basis and a starting point. Therefore there is the need – as a second step – to specify the general human rights by negotiating a human rights-based UN or UNESCO soft law declaration on "AI Ethics and Human Rights". This new declaration could and should avoid the disadvantages of the 2019 OECD AI re-

114 AI-governance means in many cases the governance of risks, as many impacts of AI are unclear and it is even unclear whether there will be something like AGI or a singularity, see above note 42. But human rights can be used as a basis for human-centered risk governance. It was *Robert Nozick* who showed that an extension of a rights-based moral theory to indeterministic cases is possible as a duty not to harm other people can be extended to a duty not to perform actions that increase their risk of being harmed. See *Silja Voeneky*, Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks, in *Silja Voeneky*/*Gerald Neuman* (eds.), Human Rights, Democracy, and Legitimacy in Times of Disorder, 2018, 153.

115 It is an obligation to protect, not only an obligation to respect; see U.N. Commission on Human Rights, Res. 2005/69, 29.04.2005, U.N. Doc. E/CN.4/2005/L.10/Add.17; Committee on Economic, Social and Cultural Rights, General Comment No 13, para. 46 (1999), reprinted in U.N. Doc. HRI/GEN/1/Rev.9, 72 (2008).

116 For the right to life, art. 6 (1) ICCPR, the second sentence provides that the right to life "shall be protected by law." In addition, the right to life is the precondition for the exercise of any other human right, part of customary international law and enshrined in all major general human rights conventions. The European Court of Human Rights has stressed the positive obligation to protect human life in several decisions; for an overview see *Niels Petersen*, Life, Right to, International Protection, in *Rüdiger Wolfrum* (ed.), Max Planck Encyclopedia of Public International Law, 2012, Vol. 6, 866. Nevertheless, the U.S. has not accepted that there exists a duty to protect against private interference due to art. 6 ICCPR; see Observations of the United States of America On the Human Rights Committee's Draft General Comment No. 36, On Article 6 – Right to Life, para. 30–38 (06.10.2017), available at: http://www.ohchr.org/EN/HRBodies/CCPR/Pages/GC36-Article6Righttolife.aspx.

117 An exception – as part of a soft law declaration – is art. 2 (b) of the Cairo Declaration on Human Rights in Islam, 05.08.1990, adopted by Organization of the Islamic Conference Res. No. 49/19-P (1990).

commendations. For this, we should identify those areas of AI-research, -development, and -deployment, which entail severe risks for core human rights.[118] A future universal "AI Ethics and Human Rights"[119] declaration should include sector-specific rules based on human rights that protect the most vulnerable rights and human dignity at the international level – as for instance, by protecting brain data. And this declaration could and should merge principles of "AI ethics",[120] as the pinciples of fairness, accountability, explainability and transparency,[121] with human rights as long as principles of AI ethics are coherent with and specify human rights in the field of AI.[122]

Silja Vöneky ist Professorin an der Albert-Ludwigs-Universität Freiburg und Direktorin des Lehrstuhls für Völkerrecht, Rechtsvergleichung und Rechtsethik sowie Fellow am FRIAS Saltus Gruppe Responsible AI.

118 I rely on those human rights that are part of the human rights treaties; whether there is the need for new human rights in the time of AI, as for a right of digital autonomy (digitale Selbstbestimmung) as the German Datenethikkommission (cf. Gutachten der Datenethikkommission, 2019, available at: https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_DE.pdf?__blob=publicationFile&v=5) argues or whether a new human right, that could be claimed by corporations, will undermine basic human rights of natural persons is still open to discussion.

119 Similar to the UNESCO Declaration on „Bioethics and Human Rights", 19.10.2005, available at: http://portal.unesco.org/en/ev.php-URL_ID=31058&URL_DO=DO_TOPIC&URL_SECTION=201.html .

120 As was shown in Part II at least some of the principles are already part of AI sector-specific regula-tion.

121 For the notion of and the need for transparency see Gutachten der Datenethikkommission, 2019, 169 et seq., 175, 185, 215 (Transparenz, Erklärbarkeit und Nachvollziehbarkeit).

122 Besides, there is the urgent need with regard to risks related to AI systems to have proactive pre-ventive regulation in place, which is backed by meaningful rules for operator incentives to reduce risks beyond pure operator liability; for a proposal see a paper by *Thorsten Schmidt*/*Silja Vöneky* on "How to regulate disruptive technologies?" (forthcoming 2020).