Table of contents

Legal Opinion

Artificial Intelligence, Responsibility, and European Regulation

Abstract

This legal opinion reflects the European approaches to foster responsible AI systems. It illuminates the potential of AI systems and what risks are inherent to it. It critically discusses the current European legislation and guidance, which aims to protect stakeholders from the harm AI can cause. Finally, the opinion articulates crucial aspects of responsible AI.

Introduction

Artificial Intelligence (AI) is said to be the next big thing. Distinguished experts such as Sir Stephen Hawking have argued that intelligent machines fuelled with big data would start a new industrial revolution. Allow this comparison to sink in for a moment. The industrial revolution is the singularity in the history of humanity. It is the very moment that let the thousands of years stagnant productivity of the world quadruple within a century – only to multiply it by 30 times within the next century. This should give us an idea of what impact we may have to expect from AI.

In countless publications, AI enthusiasts and naysayers are quarrelling about its prospects and risks. But as Mahatma Gandhi said, honest differences are a healthy sign of progress. And both camps are on the same page in expecting AI is here to stay; for better or for worse. Assuming that it is in the interest of AI developers to avoid their stakeholders suffering harm, approaches for responsible AI are in high demand.

This work will shine a light on the current governance side of this technology. In *Chapter 1*, we jump in at the deep end and explore what AI, machine learning and neuro networks are, what the technology is capable of, and which risks are inherent to it. *Chapter 2* presents and scrutinises European legislation and guidance that has been issued to date to protect AI stakeholders and carves out common features and differences. On that basis, *Chapter 3* critically discusses the illuminated attempts and identifies facets crucial for the development and usage of responsible AI.

Chapter 1 – Artificial Intelligence, its risks, and stakeholders

In 1955 John McCarthy coined AI as a term for the concept of mimicking human intelligence with machines.[1] He argued that intellectual differences between human and machines (artefacts) were illusory and a matter of time to solve.[2] According to him, intelligence is demonstrated by an artefact if it (1) evidences perception and cognition of relevant aspects of its environment, (2) has goals, and (3) formulates actions towards the achievement of these goals.[3] This chapter scrutinises AI's technical possibilities (see under I.), its risks (see under II.), and the stakeholders and shareholders of AI applications (see under III.).

I. AI's technical possibilities and limits

Later, AI became the umbrella term for an array of intertwined technologies from which machine learning is the current cornerstone. Machine learning made AI far more potent than in the past. It is the ability to adapt to an environment (training data) without human intervention (supervised or unsupervised).[4] Machine learning algorithms can be trained whereby they develop specific behavioural patterns in which's process the algorithm may rewrite itself. The most current example of machine learning is artificial neural networks. They differ from ordinary computer processing by imitating actual human neural networks.[5]

Only the recent leap of computer processing capabilities made machine learning and artificial neural networks feasible. When artificial neural networks are trained with data (and thereby given a task), they are not acting deterministically and, therefore, predictably like traditional algorithms. Instead, they deal with the environment probabilistically and empirically by practising different behaviour. That leaves both future predictions and past explanations of the behaviour with a certain degree of uncertainty.[6] Today, artificial neural

[1] *McCarthy* et al., 1955.; The concept reaches back to *Alan Turing* who in 1950 proposed to consider the question: "Can machines think?": *Turing*, Mind 1950, 433; *Turing's* question reaches back to *Descartes* who asked whether machines can imitate the behaviour and thinking of living beings, *Descartes*, Discourse 1637, 44

[2] A review of the discussion of that time can be found at *Campolo/Crawford*, E.S.T.S. 2020, 2 f.

[3] *Clarke*, C.L. & S.R. 2019, 423, 424, referring to *McCarthy*, 2007.

[4] *Zech*, GRUR Int. 2019, 1145 f.

[5] *Clarke*, C.L. & S.R. 2019, 423, 425.; *Zech*, GRUR Int. 2019, 1145 f.; further explanation and illustration can be found at: House of Lords, AI Report 2017, 15 f.

[6] *Zech*, GRUR Int. 2019, 1145 f.; *Clarke*, C.L. & S.R. 2019, 423, 425 f.; *Campolo/Crawford*, E.S.T.S. 2020, 8 f.

networks are pushing boundaries of complexity, reaching an extent of up to 10^8 "simulated neurons" connected with up to 10^11 "simulated synapses".[7] Correspondingly, the complexity of tasks AI can perform is progressing and more and more reaching into human domains.

AI means enabling machines to perform tasks that are being considered to be human domains in four areas, namely, learning, speech, vision, and language. As of now, AI has been used to decipher and reproduce human voice inputs successfully and defeated human world champions in strategic game systems like chess, Go, and computer games. AI composes deceptive photos and videos and writes books and articles. It is also being used in driving assistance systems, high-frequency trading and cancer diagnosing.[8] With investments to be made and continuing rises in computational power, it is to be assumed that we are just scratching the surface of what it can deliver.[9] AI is expected to help reduce human impact on the environment and improve the efficient use of natural resources and energy, improve energy infrastructure and consumption and help ameliorate transport systems, advance health care with better diagnosis and treatment, and help social forecasting, e.g. regarding educational and professional opportunities.[10]

At this day and age, Google's "Inception" artificial neural network performs image recognition unsupervised. It recognises pictures on the basis of its training data (i.e. prior labelled pictures). The recognition of a specific picture is a probabilistic decision in which Inception matches the features of the specific picture with its training data. It does that stunningly well. However, Inception cannot provide an explanation for its decisions.[11] Moreover, relatively small distortions to inputs can create massive increases in its error classification rates. MIT students altered the pixels of a cat picture (which a human would immediately identify as such) so that Inception recognised it with maximum probability as a picture showing guacamole. The doubtless ironical notion vanished with Inception's second acid test: After repainting a 3D-printed turtle (still immediately identifiable for a human),

[7] *Zech*, GRUR Int. 2019, 1145 f.

[8] These and more examples given by Arruda, A.J.T.A. 2017, 448 ff.

[9] In macroeconomic terms, it is expected that AI would boost the overall level of economic activity, the pace of economic growth, and will result in a better balance between work and leisure, see *Bootle*, 2019, 70, 83 ff., 97 ff.

[10] EU AI Guidelines, 32 f.

[11] *Flett/Wilson*, C.T.L.R. 2017, 72 f., *Whittaker*, J.I.B.L.R. 2019, 296.

Inception's classified it as a rifle.[12]

II. Sectorial risks of AI applications

The illustration shadows forth what can be at stake when AI algorithms interpret their environment, have goals and make automated decisions with regards to it. The risks inherent to AI can be assigned to the different points in time of the technology's application: conceptual phase, deploying, and output. During the conceptual phase, before an AI tool is deployed, the errors and biases of operators may manifest in setting the environment for the tool's usage: The conclusions drawn by AI will reflect the errors and biases indwelling in the implicit model, the chosen learning algorithm, the selection of training data, and the selection of real-world circumstances for which the training data was created.[13] After the AI tool has been deployed, during its performance, risks inherent to AI's data reliant inferencing process may emerge: The data used by AI may be of insufficient quality *(data quality*),[14] the AI may make inappropriate assumptions about data or make inappropriate assumptions about the inferencing process itself (*process quality*).[15] Finally, when looking at the AI's output, risks and flaws of automated probabilistic decision-making become apparent: The AI's inferencing process is opaque in nature and conclusions drawn may not be traceable (*transparency*),[16] nonetheless the AI may have autonomously carried out decisions with regards to it (*autonomy*), and, consequently, responsibilities for decisions made and harm caused by an AI may be difficult to assign due to the complexity of the AI system and its supply chain (*accountability*).[17]

These risks loom even larger considering the proposed scale of AI usage and the

[12] See *Zittrain*, 2019 for the full story (and pictures of the deceptive animals); more examples are given by *Campolo/Crawford*, E.S.T.S. 2020, 8 ff.

[13] *Clarke*, C.L. & S.R. 2019, 423, 426; *Yu/Ali*, L.I.M. 2019, 3 f.; Google's photo app horribly misclassified pictures of black people as gorillas, reflecting the biased training data given to it that mainly consisted of white people, see House of Commons, AI Report 2016, 18.

[14] That applies particularly to AI systems that draw data from multiple sources where quality consistency is difficult to maintain; the UK Government addressed the general issue of data quality for data intensive projects in its Data Ethics Framework 2018.

[15] *Clarke*, C.L. & S.R. 2019, 423, 426 ff.; *Coeckelbergh*, S.E.E. 2019, 12, 14; *Yu/Ali*, L.I.M. 2019, 4 f.

[16] *Kemp*, Comms.L. 2019, 32f.; Google's AI Alpha Go beat the world champion in Go with a highly unusual move that stunned commentators so that they assumed the AI had malfunctioned, it could not be explained why it made that move and what its rationale was, see: House of Commons, AI Report 2016, 17.

[17] *Clarke*, C.L. & S.R. 2019, 423, 428 f.; *Yu/Ali*, L.I.M. 2019, 5 f.; *Whittaker*, J.I.B.L.R. 2019, 296 f., the global supply chain for AI systems can further hamper litigation as it makes obtaining necessary information about the AI's design, coding, etc. extra difficult, see: House of Lords, AI Report 2017, 96.

current ownership structure of the necessary means to deploy AI tools. AI applications are designed to operate on a large scale and high speed in sensitive areas whether it be pricing insurances, diagnosing tumours, autonomous driving or even military actions.[18] Depending on the application, both its merits and harm will have significant effects on numerous individuals or organisations. The harmful potential is further fuelled by the ever-increasing persuasiveness of the data allocated. Moreover, the kind and volume of data necessary for producing useful predictions is likely to be in the hands of Google, Amazon, and Facebook rather than the public.[19] When we, finally, think about crime in liaison with autonomous cars or weaponry, drones, and the connected economy, the scope for mischief seems amplified.[20] A series of questionable applications are already in train.[21]

III. Stakeholders and shareholders of AI applications

AI's inherent risks are multifactorial and concern shareholders and stakeholders of AI applications alike. Shareholders are sensitive to any risks affecting the applicator's vital interests. In case the applicator is an entity, its shareholders suffer from (potential) decreases in the company's business prospects and its asset value. Inadequate use or usage of malicious AI could, for instance, directly harm the company's assets and impair its prospects by giving the reason for legal claims of customers or provoke bad press. AI's misuse or malfunction could also affect the vital interests of applicators in the public sector. Its direct shareholders are high-profile political and administrative personnel whose careers are tightly connected to a implementing a political agenda. Again, inadequate use or implementing malicious AI for that means could easily backfire against the shareholders' vital interests.

Stakeholders, on the other hand, are exposed to a wide array of risks. According to a broad definition in the context of information technology, stakeholders are not just users but also "any other individuals, groups, or organisations whose actions can influence or be influenced by the development and use of the system whether directly or indirectly."[22] Direct

[18] *Coeckelbergh*, S.E.E. 2019, 14; *Whittaker*, J.I.B.L.R. 2019, 296.

[19] *Zittrain*, 2019.

[20] *Rowe*, J.P.I.L 2018, 307; *Yu/Ali*, L.I.M. 2019, 5.

[21] Such as identifying and tracking of individuals, usage of facial recognition and biometric data, lie detection, personality assessment by means of micro-expressions or voice detection, disguising AI systems so that individuals may not grasp they are dealing with a machine, citizen and customer scoring, particularly, assessing "ethical integrity" and "moral personality", and finally, autonomous weapons, see: EU AI Guidelines, 33 f.

[22] *Clarke*, *Clarke*, C.L. & S.R. 2019, 410, 413 (citing: *Pouloudi/Whitley*, 1997, 3).

impact could emerge due to biased datasets or algorithms impairing impartial classification of persons using AI, e.g. for credit scoring or job applications.[23] Malicious inferences drawn from sensors of self-driving cars misclassifying objects can lead to direct harm of the AI-using driver as well as other road users.[24] Another virulent threat are malfunctioning facial recognition systems for policing with wrongly targeted law enforcement actions.[25] Stakeholders may also get impaired indirectly through social or economic profiling or scoring of parts of the population or certain areas and political or administrative consequences such as increased police attention, the closing of welfare facilities, decrease in public funding.[26] That sketch of AI threats can be extended at will. Consequently, the risks AI poses may loom as large as its impact on business and society.

Chapter 2 – European attempts of addressing the risks inherent to AI

René Descartes rejected the idea that machines can be subject to reason because they "do not act with sense but only according to the disposition of their organs".[27] Likewise, technology is per se neutral and indifferent to law and ethics. Only its application complements or contradicts societal principles. Fostering and reaping the fruits of technology is therefore subject to regulation and other tools formulating norms for appropriate conduct.[28] Given the early stage at which we are in the application of AI systems, actual legislation addressing its risks and forming best-practices has been marginal to date. The sole beacon shining a light on AI in the field of hard law is the EU General Data Protection Regulation (GDPR) (see under I.). On the other hand, a plethora of guidance has been released from numerous institutional bodies with the aim to influence the formation of industry standards. The EU issued its Ethics Guidelines for Trustworthy AI (see under II.), the UK published its guide to using AI in the public sector (see under III.). These shall be illuminated on the

[23] *Yu/Ali*, L.I.M. 2019, 5 reflects research according to which employers are 50% more eager to hire applicants with white-sounding names than those whose names are African-sounding. Also, reference is made to Amazon's hiring AI that discriminated female candidates.

[24] House of Lords, AI Report 2017, 95 f.

[25] Only recently and in the lights of the protests, lootings, and policing in reaction to the death of George Floyd in the U.S., tech firms announced that they refrain from providing the police with facial recognition systems before it is governed by a national law "grounded in human rights", see Washington Post, June 11, 2020.

[26] A comprehensive listing of AI misconduct can be found at *Yu/Ali*, L.I.M. 2019, 3 ff.; more examples given by *McGregor et al.*, I.C.L.Q. 2019, 68(2), 315 f.

[27] *Descartes*, Discourse 1637, 44.

[28] Laws and conduct guidelines are often released in reaction to inventions and their implementation in daily life, such as road safety laws, product liability laws, consumer laws, or only recently, the GDPR for processing personal data.

principles they emphasise and how they aim to mitigate AI's threats for stakeholders.

I. Art 13(2)(f) and 22 GDPR

The GDPR recognises the effects of algorithmic decision-making on fundamental personal rights and freedom of natural persons. Art 13(2)(f) GDPR imposes an obligation on controllers of such processing to inform data subjects when personal data is obtained about "the existence of automated decision-making, including profiling" and provide meaningful information about the logic involved, the significance as well as envisaged consequences of such processing. According to Art 22(1) GDPR data subjects shall not be subject to a decision based solely on automated processing, including profiling, which produces legal effects, particularly when processing sensible special categories data of Art 9(1) GDPR such as race, political opinion, gender, etc. Recital 71(2) makes it clear that this particularly includes decisions made by AI systems such as to "analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, […]".[29] If permitted, [30] such processing may only be carried out if the data subject retains the right to obtain human intervention, to express his point of view, obtain an explanation of the decision and challenge it ("human in the loop"), Recital 71(4).[31]

The provisions seek to promote transparency and accountability. Transparency is sought by informing the data subject about the fact and the underlying logic of the automated decisions made with respect to it. Accountability for such processing is achieved through obligating the controller as the beneficiary of the processing to inform data subjects about the means as well as obligating them to ensure the data subjects can exercise their rights to human intervention and challenging the decision. Yu and Ali argue that these provisions would prohibit automated decisions of AI systems and prohibit utilising algorithms that are taking into account variables such as gender, race, or religion.[32] In the author's opinion, however, this interpretation is not convincing. Instead, the provisions should rather be seen as borders and instructions to such intrusive measures. Therefore, Recital 71(6) demands controllers to justify such intrusive measures by ensuring "fair and transparent processing"

[29] According to Recital 71 (1) that, e.g., includes automatic refusal of an online credit application or e-recruiting practices without human intervention.

[30] Such processing is permittable for a contract between the data subject and the controller if suitable regulatory safeguards are in place, or the subject's consent has been obtained, Art 22(4) GDPR.

[31] So-called "right to explanation": *Flett/Wilson*, C.T.L.R. 2017, 73.

[32] *Yu/Ali*, L.I.M. 2019, 3.

through the implementation of "technical and organisational measures" preventing "discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, […]", such as "data sanitisation".

Data sanitisation addresses data quality and biases by requiring the eliminating of potentially biasing categories variables of the special categories data of Art 9(1) GDPR. Data sanitisation includes potentially explicit discriminatory variables such as gender as well as potentially implicit discriminatory variables such as the weight or height to distinguish men and women.[33] Data sanitisation emphasises the importance of data quality for unbiased automated processing of data, including by means of AI systems. It requires implementers to scrutinise their goals and envisaged means by taking into account how they might affect the legitimate interests of affected data subjects. This assessment of proportionality reaches far into the conceptual stage of the implementation of an automated processing system requiring the implementer to weigh their interest in automating decisions on the basis of certain variables against the legitimate interest of the affected data subjects for an unbiased decision.

It should be noted that the GDPR is only first step into regulating AI usage with stakeholder's rights and interests in mind. This observation applies even more given the fact that it is the first codified law that actually obligates operators of AI to meeting standards safeguarding the rights and legitimate interests of data subjects affected by the processing. The principles of data quality, transparency, and accountability are promoted. Time will tell whether the GDPR will apply in the UK after the Brexit transition period. Given the gold-standard reputation the GDPR received and its blueprint-role for other legislation, it may well be that its approach on AI and automated decision-making will also be echoed in the future.

II. EU Ethics Guidelines for Trustworthy AI

The EU Ethics Guidelines for Trustworthy AI (EU AI Guidelines) are one constituent part of what the European Commission pictures as an "ethical, secure and cutting-edge AI made in Europe".[34] That architecture rests on three columns, namely, increasing public and private investments, preparation for socio-economic change, and developing and implementing a capable ethical and legal framework.[35]

[33] *Yu/Ali*, L.I.M. 2019, 4.

[34] EU AI Guidelines, 4.

[35] EU AI Guidelines, 4.

Legally, the Guidelines are fundamentally different from the GDPR. Rather than focussing on sector-specific requirements such as processing personal data as the field of application of the GDPR, they follow a general approach in addressing stakeholder-centric AI issues across public and private sectors.[36] And most importantly, they are no legislation and do not contain binding rules. Instead, the EU AI Guidelines are developing the idea of a "Trustworthy AI" by addressing ethical principles concerned, presenting a set of requirements on their premises, and proposing a procedure for compliance.

Such Trustworthy AI consists of three components which should be met throughout an AI system's entire life cycle: *lawfulness* (complying with all applicable laws and regulations); *ethicality* (ensuring adherence to ethical principles and values); and *robustness* (both technically and socially).[37] The framework for Trustworthy AI itself follows the AI-life cycle by, firstly, adhering to the ethical principles based on fundamental rights, secondly, implementing the key requirements derived, and thirdly, operationalising the key requirements to specific AI applications.

1) Legal Foundations of Trustworthy AI

Notably, the underpinning legal obligations for lawful AI are not elaborated upon. The Guidelines are settling for it by sketching the grounds for lawful AI as follows: EU primary law which includes the Treaties of the European Union (EU Treaties) and its Charter of Fundamental Rights (EU Charter). EU secondary law such as the GDPR, the Product Liability Directive (Directive 85/374/EEC), the Regulation of the Free Flow of Non-Personal-Data (Regulation 2018/1807) and other EU-supranational provisions. And, member states law as well as sector-specific rules and obligations.[38]

The ethical framework is not newly invented but said to be derived from existing "fundamental rights", including the EU Treaties, the EU Charter, and international human rights law.[39] From the EU Treaties, Art 2 and Art 3 are identified which protect the fundamental and indivisible rights of human beings and guarantee the rule of law, foster

[36] Still, the EU AI Guidelines are recognising that specific sectors are likely to become subject for AI-specific legal rules, see EU AI Guidelines, 6; sectors that come to mind are FinTech, LegalTech, BioTech, MedTech, etc.

[37] EU AI Guidelines, 5, 6 f.

[38] EU AI Guidelines, 5.

[39] EU AI Guidelines, 9.

democratic freedom and promote the common good. The Guidelines do also not go into detail on the EU Charter rights concerned, but make a short reference to “dignity, freedom, equality and solidarity, citizen’s rights and justice”.[40] All have a common denominator: The respect for human dignity which is called a “human-centric approach”. That approach grants the human being a unique and unalienable moral primacy in the civil, political and social fields.[41]

The Guidelines do not elaborate on the human rights law basis. From international human rights law, the International Covenant on Civil and Political Rights (1966, ICCPR) come to mind as potential legal grounds. According to Art 2(1) and (2) ICCPR, states must prevent human rights violations through setting up a framework for monitoring and oversight measures, promoting accountability, and ensuring remedy to individuals.[42] These obligations do not only apply directly to state actions. Instead, there is an indirect application on violations from third parties, including businesses, which the state guarantees to prevent.[43] The Office of the High Commissioner for Human Rights emphasises that Art 2(1) ICCPR requires states to set up all legislative and jurisdictional measures necessary to ensure “appropriate” protection.[44] The ICCPR is codified by EU member states, the UK, and U.S., the relevant countries for this work.

This culmination of international law could be seen as legal grounds for the EU AI Guidelines and future legislation. McGregor et al. argue that the recurred states’ obligation to provide for a framework preventing human rights violations should include governing AI systems for their harmful potential.[45] And indeed, the above-sketched risks of predictive AI systems to the legitimate interests of shareholders and stakeholders are capable of violating human rights. For instance, AI enhanced policing or social scoring could interfere with the right to liberty, the prohibition of discrimination, and the right to privacy. The same applies to other forms of AI-enhanced public or private predictions and judgements such as automated pricing for insurances and risk assessments for domestic abuse. These looming threats to

[40] EU AI Guidelines, 10.

[41] EU AI Guidelines, 10.

[42] UN HRC, Comment No. 31, paras 3-7.

[43] UN HRC, Comment No. 31, para 8; UN HRC, Ruggie Principles, Principles 1-10; the Ruggie principles further formulate expectations that businesses should respect human rights in the way that they should avoid infringing human rights of others and should address adverse human rights impacts with which they are involved, see Principle 11.

[44] CESCR Comment No. 3, paras 2-9.

[45] *McGregor et al.*, I.C.L.Q. 2019, 68(2), 325 ff.

human rights are intensified by the black-box architecture of current AI systems that impair their predictions' and decisions' accountability and redeemability.

On the other hand, McGregor's justification of Art 2(1) and (2) ICCPR as the legal basis for ethical AI regulations raises questions. The ICCPR ruleset focusses on the protection of human rights violations. As such, AI systems to date, however, remain a to be developed technology with a limited market and public sector readiness. In my opinion, their lack of readiness and to be explored nature do not yet justify the risk analysis for human rights violations that are needed in order to implement concrete state regulations. Clearly, the risk analysis can be undertaken with more acceptance when narrowing down the sector of AI systems' application. In certain fields say autonomous driving or weaponry, applications have already emerged, allowing for sufficiently accurate risk analysis and subsequent regulations.[46] In other fields say entertainment or retail customer relationships, AI applications are only slowly emerging and still lack the distinctiveness, scale, and human rights violation potential in order to justify preventive regulations. Consequently, McGregor appears to jump to conclusions when demanding preventive regulation for the entirety of AI systems on the basis of the ICCPR rules. In the author's view, Art 2(1) and (2) ICCPR demand a comprehensive and ongoing risk analysis of potential human rights violations by different AI applications before obligating the underwriters to implement additional human rights protecting regulations.

However, the EU AI Guidelines expressively point out that they do not focus on the lawful side of AI.[47] This may be the reason why the Guidelines go no further into detail about their legal foundations. Yet, there is still a compelling reason for placing the Guidelines on the firm foundation of the EU Charter and human rights. Resting them on these grounds allows claiming some legitimacy for its principles despite the legally unbinding nature.[48] That impression is further complemented with the Guidelines focus on practical procedures for AI utilisation rather than overall societal implications, giving the EU AI Guidelines the complexion of a compliance-manual for operators. That design could, ultimately, increase the chances that the Guidelines' principles and procedures find incorporation in AI systems.

[46] Equally, one could pinpoint certain ways of processing data through AI systems that bear distinctive risks for human rights as, for instance, with the GDPR rules, see above under Chapter 2, I.

[47] EU AI Guidelines, 6.

[48] *Whittaker*, J.I.B.L.R. 2019, 297.

2) *Principles of Trustworthy AI*

To foster fundamental values, the Guidelines identify four “ethical principles” essential for Trustworthy AI, namely, the respect for *human autonomy*, the *prevention of harm*, *fairness,* and the principle of *explicability*.[49] On that basis, seven equally important requirements for Trustworthy AI are set out:

1. Human agency and oversight (including fundamental rights),
2. Technical robustness and safety (including resilience to attack and security and general safety, accuracy, reliability and reproducibility),
3. Privacy and data governance (including respect for privacy, quality and integrity of data, and access to it),
4. Transparency (including traceability, explainability, and communication),
5. Diversity, non-discrimination and fairness (including avoidance of unfair bias, accessibility and universal design, and stakeholder participation),
6. Societal and environmental well-being (including sustainability and environmental friendliness, social impact, society and democracy),
7. Accountability (including auditability, minimisation and reporting of negative impact, trade-offs and redress).[50]

From the requirements for Trustworthy AI the following aspects leap out at us. Firstly, the requirements breathe the air of AI systems with the purpose to supplement human life and strive rather than replacing it. That is best embodied by requirement 1 whose explanation formulates that AI systems are supposed to “support human autonomy and decision-making” and demands them to “act as enablers to a democratic flourishing and equitable society by supporting the user’s agency and foster fundamental rights”.[51] The emphasis on the human control of AI systems’ predictions and decisions shows parallels to the right to human intervention of Art 22(2) GDPR and Recital 71(4).[52] However, it does not simply require a human in the loop, i.e., the capability for human intervention in every

[49] EU AI Guidelines, 12.

[50] EU AI Guidelines, 14.

[51] EU AI Guidelines, 15.

[52] See above under Chapter 2, I.

decision of the system but goes further. Depending on the system, it may be necessary to have a "human on the loop", i.e., the ability to intervention in the design and operation of AI systems, or a "human in command", i.e., the ability to overwatch the system and to decide the if and how of its application.[53] AI operators are requested to tie the level of human control to the reasoning of proportionality where a minus in oversight must be mirrored with a plus in prior testing and governance. This demand goes beyond Art 22(3) GDPR, whose risk assessment circulates around the right to privacy. Instead, the EU AI Guidelines require an exhaustive risk assessment with respect to the autonomy of AI systems and their trade-offs.

The emphasis on human supplementation is complemented by the technical robustness and safety requirement that focusses on the prevention of harm but also requires explicability as AI system's output should be reproducible. The right to privacy should be guaranteed throughout the AI system's life cycle, including not only information provided by the user but such information generated about the user over the course of their interaction with the system. That is, again, mirrored by Art 22(3) GDPR and Recital 71(4) and the GDPR's human-centric focus. Moreover, the EU AI Guidelines also focus on data quality and integrity and the need for data sanitisation. Processes and data sets should be tested and documented at each step of the AI life cycle.[54] The transparency requirements also require AI systems to inform users about their nature and not to represent themselves as humans. As in the case of Art 22(3) GDPR, users should be given an option to decide against the AI in favour of human interaction.[55]

Through the requirements of diversity, non-discrimination and fairness, the conceptual stage of AI usage is also being addressed. As the conclusions drawn by AI will reflect errors and biases indwelling in the utilised model, learning algorithm, training data, and the chosen circumstances of data collection, [56] all stages of the conceptual side of AI application are addressed. Consequently, "identifiable and discriminatory biases should be removed in the collection phase" and AI systems themselves should be kept unbiased by monitoring and "hiring from diverse backgrounds, cultures, and disciplines".[57] AI controllers should not only guarantee diversity for their developers and monitors but are expected to

[53] EU AI Guidelines, 16.

[54] EU AI Guidelines, 17.

[55] EU AI Guidelines, 19.

[56] *Clarke*, C.L. & S.R. 2019, 423, 426; *Yu/Ali*, L.I.M. 2019, 3 f.

[57] EU AI Guidelines, 18.

foster “universal design” principles so that users with diverse abilities and disabilities are enabled to access the AI applications.[58]

The EU AI Guidelines go beyond the technical and conceptual stage of AI systems by addressing societal and environmental well-being with requirement 6. The Guidelines claim that “broader society, including other sentient beings and the environment, should also be considered as stakeholders” that are to be considered and protected when bringing AI applications forward.[59] Responsible AI should not just be focussed on stakeholders as “any other individuals, groups, or organisations whose actions can influence or be influenced by the development and use of the system whether directly or indirectly”[60] but also consider its ecological and sustainable footprint. Therefore, it should be ensured that AI systems work in the most environmentally friendly way possible, applying to its development, deployment, and use processes as well as its entire supply chain.[61]

The beforementioned principles are safeguarded by the accountability requirement. Its emphasis on auditability and constant internal and external monitoring of all stages of AI systems is seen as key in identifying risks and maintaining compliance throughout AI operations.[62] Impact assessments should be conducted before and during the operation, for instance, with the use of “red-teaming”[63] including ethical hackers and algorithmic assessments. Furthermore, whistle-blowers, NGOs, trade unions and other entities reporting legitimate concerns about AI systems should enjoy protection. The principle of proportionality should guide AI controllers throughout their operations. Where trade-offs occur in the system’s life-circle with they should be “explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights”.[64] Moreover, where no ethically acceptable trade-off can be found, the “development, deployment and use of the AI system should not proceed in that form.”[65] The decision-maker must be accountable for the trade-offs being made and should review the appropriateness regularly so that adjustments

[58] EU AI Guidelines, 18 f.

[59] EU AI Guidelines, 19.

[60] See above under Chapter 1, III.

[61] EU AI Guidelines, 19.

[62] EU AI Guidelines, 19 f.

[63] A practise whereby an independent unit challenges an organisation from an adversarial role to improve its effectiveness, particularly, to help identify security leaks, EU AI Guidelines, 37.

[64] EU AI Guidelines, 20.

[65] EU AI Guidelines, 20.

can be undertaken where required.

3) *Implementation of Trustworthy AI*

The EU AI Guidelines offer a toolbox of measures to implement the principles of trustworthiness in actual AI systems. AI systems should include the requirements from scratch so that their architecture becomes "ethical by design".[66] This should be accomplished by technical and organisational measures. Technical measures emphasise setting up "white list" rules concerning states and behaviours which the system should follow and "blacklist" restrictions of such actions the system should never surpass.[67] AI systems should include mechanisms for fail-safe shutdowns and enable resumed operations after forced shutdowns, and the "sense-plan-act" cycle[68] should be adapted for AI systems. Compliance should be sought through constant monitoring and testing. It should go beyond standard methods of bug testing prior and after launch but involve deliberate attempts to breaking the system by using "red teams" and offer "bug bounties" so that outsiders are incentivised to locate and report weaknesses in the AI system. Quality of service indicators for algorithms and data are seen in testing and training as well as in traditional software metrics. [69]

Organisational measures consist of encouraging AI controllers and stakeholders to sign up to the Guidelines. Its principles should be incorporated into codes of conduct, industry design standards such as ISO or IEEE, professional codes of ethics or a "Trustworthy AI" accreditation and label.[70] Organisations should appoint persons responsible for AI ethics or make us of internal or external AI ethics boards. Design teams should be hired from hiring from diverse backgrounds, cultures, and disciplines. Additionally, all stakeholders should get encouraged to informed participation through communication,

[66] *Whittaker*, J.I.B.L.R. 2019, 299; that phrase is coined from the terms "secure by design" and "private by design", for the privacy context see above under Chapter 2, I.

[67] EU AI Guidelines, 21.

[68] Three-step system for limiting adversarial effects of machine doings through 1) sense (gathering information using sensors), 2) plan (develop a world with all environmental elements necessary and plan the next move), and 3) act (in the context with Trustworthy AI, only actions should be taken that are compliant with the requirements set up), see EU AI Guidelines, 21.

[69] EU AI Guidelines, 22.

[70] Interestingly, such accreditation and labelling for trustworthy or responsible AI controllers is currently brought forward by the industry as well as smaller players, see, for instance, https://ai-global.org/2020/04/28/creating-a-responsible-ai-trust-index-a-unified-assessment-to-assure-the-responsible-design-development-and-deployment-of-ai/.

education and training, e.g., in the form of discussing panels.[71]

III. UK guide to using AI in the public sector

The House of Commons' Science and Technology Committee quotes Innovate UK's view: "Appropriate legal and regulators frameworks will have to be developed to support the more widespread deployment of robots and, in particular, autonomous systems. Frameworks need to be created to establish where responsibilities lie, to ensure the safe and effective functioning of autonomous systems, and how to handle disputes in areas where no legal precedence has been set."[72] The UK Office for Artificial Intelligence picked up that pass and issued its first framework for using AI in the public sector (UK AI Guide). The UK AI Guide shows close links to the EU AI Guidelines and seeks to give practical guidance for AI developers that aim to provide AI applications to the UK public sector. This is particularly relevant for digital health care services offered to the UK's National Health Service as it is a public sector organisation that requires compliance with the UK AI Guide.[73]

According to the UK AI Guide, AI developers should (1) understand their envisaged AI system, (2) assess, plan, and manage their AI system, and, importantly, (3) use their AI systems ethically and safely.[74] The guide emphasises that applicable law must be complied with and expressively refers to the GDPR, namely Art 22 GDPR, and its relevance for automated decision-making.[75] Interestingly, the UK AI Guide also interprets Art 22 GDPR so that it only affects such automated processes that bear decisions with legal or similarly significant effects on individuals.[76] Hence, less intrusive AI decisions should also not fall under the provisions regime. The strong parallels to the EU framework continue when looking at the requirements for ethical and safe AI.

The UK AI Guide demands from AI developers to start their projects with a framework of ethical values for a responsible design of AI which (1) respects the dignity of individuals, (2) connects with others sincerely, openly and inclusively, (3) cares for the well-

[71] EU AI Guidelines, 23.

[72] House of Commons, AI Report 2016, 22.

[73] Notably, digital health is one of the most important industry sectors for the UK's AI strategy where university hospitals launched a series of projects with Google's AI firm DeepMind, see House of Commons, AI Report 2016, 34.

[74] UK AI Guide, 14 ff., 22 ff., 34 f., 38 ff.

[75] UK AI Guide, 12 f., 27.

[76] See above under Chapter 2, I.

being of all, and (4) protects the priorities of social values, justice and public interest.[77] On the basis of these values, developers should establish a set of actionable principles to consider the ethical permissibility of their AI project. These principles, too, emphasise (1) fairness, (2) accountability, (3) sustainability, and (4) transparency.[78] Fairness should be achieved through using sanitised and equitable datasets and preventing the system "from having discriminatory impact". Accountability should help making the AI system fully answerable and auditable by installing a continuous chain of responsibility and implementing activity monitoring throughout the project's lifespan. Sustainability emphasises technical robustness and longevity but also the transformative effects AI can have on individuals and society. Lastly, transparency demands explainability for "affected stakeholders" as well as justification for the ethical permissibility, discriminatory non-harm, and the public trustworthiness of the AI application's outcome and of the processes behind its design and use.[79]

The UK AI Guide's principles are clearly designed in line with EU AI Guidelines. While, in contrast to the EU Guidelines, the UK Guide's ethical values do not expressively stress the developer's responsibility for the environment and society as a whole they still seem to operate with the same broad definition of AI stakeholders as the EU AI Guidance, when identifying the task to "care for the wellbeing of all".[80] The abstract and broad-stroke formulation of the UK's AI principles and the framework that sketches those terms only with vague descriptions are open for interpretation and should also only be seen as a starting point for organisations who consider operating with AI systems. Clearly, the Guide does lack the specific step-by-step schedule for AI developers which the EU AI Guidelines provide. But it offers a first overview of the principles to be integrated in process-based governance frameworks that address the relevant persons and roles involved, the relevant workflow and its stages, frames for evaluations and monitoring, as well as considerations about the ethical and legal interests concerned when applying AI systems.

Chapter 3 – Analysis, discussion and suggestions for fostering responsible AI

The EU AI Guidelines and the UK AI Guide offer a comprehensive first step to shape and regulate AI systems while the EU GDPR contains the first AI specific legal provisions.

[77] UK AI Guide, 40.

[78] UK AI Guide, 42.

[79] UK AI Guide, 42.

[80] UK AI Guide, 40; for the discussion around the definition of stakeholder under the EU AI Guidance see above under Chapter 2, II. 2).

The following aspects of the European attempts for encouraging responsible AI deserve further attention (see under I. 1) – 5)). Against their light, one can formulate themes and principles crucial for developing responsible AI (see under II.).

I. Analysis and Discussion

The following aspects of the European AI guidelines to date appear controversial.

1) Vague principles and the burden of uncertain trade-offs

In the author's opinion, the broad strokes of the guidelines' definitions and application bear challengers. The biggest challenge appears to be the vagueness of its approach and scope of application.[81] The EU AI Guidelines expand under in the introduction to chapter 2 and in its requirements 2 and 6 the definition of stakeholder significantly. According to the explanation of technical robustness and safety, it must be ensured that the system will do what it is supposed to do "without harming livings beings or the environment".[82] In the same manner, the requirement for societal and environmental well-being includes "broader society, including other sentient beings and the environment" as stakeholders.[83] In the same manner, the UK AI Guide demands the "care for the wellbeing of all".[84] This scope goes far beyond the broad definition for shareholders set out above.[85] The scope of application does not only affect direct users and such individuals that are affected by the specific usage of AI systems but society as a whole, all living things, and the environment. It is understandable and agreeable that the guidelines emphasise the responsibility of AI developers for taking actions in society and natural ecosystems. A technology which is deemed to transform both should not be taken lightly and, instead, be developed and applied so that its collaterals and harm created are limited to the minimum. However, sketching the scope so limitless that it should protect everything and everyone while not focussing on certain target groups, rights, or causes for harm exposes the guidelines to the danger of meaning everything and nothing. This holistic approach and the absence of a clear focus of protection could result in two short fallings.

[81] See also *Salami*, C.T.L.R. 2020, 127; *Whittaker*, J.I.B.L.R. 2019, 299; *Clarke,* C.L. & S.R. 2019, 410, 415.

[82] EU AI Guidelines, 17.

[83] EU AI Guidelines, 19.

[84] UK AI Guide, 40.

[85] See above under Chapter 1, III.

Firstly, it could result in a lack of protection of vital interests such as human rights of stakeholders in the meaning of direct and indirect participants of AI systems because they get lost in the shuffle of different scopes and premises. For instance, designing an AI application so that it is most energy-efficient and holds the best ecological footprint does by not have any necessary implications for protecting the human rights or privacy of its stakeholders or encouraging transparency. Even the contrary could be imaginable when trade-offs are made in favour of saving energy so that, for instance, back-ups of decisions made by the AI are saved only for a short period, limiting their explainability and transparency. The other side of the coin could be infinite storage of data processed and decisions made by the AI and access guaranteed to stakeholders at all times which would, in contrast, cause much higher energy consumption and increase its ecological footprint. While trade-offs like the abovementioned will become routine in developing and applying AI systems, the guidelines do only offer limited assistance for making such decisions. Its emphasis on fundamental rights when trade-offs have to be made does contain some indication towards the stronger valuation of human rights when in conflict with say sustainability and environmental friendliness.[86] Still, developers are supposed to be held fully responsible and the EU AI Guidelines ultimately demand termination of the AI system where “no ethically acceptable trade-off” can be found. In that light, trade-offs between the multitude of interconnected factors are likely to pose dilemma situations on decision-makers.

This holds even more potential for uncertainty as some requirements are double-edged and seem to contradict each other. For instance, designing AI applications with maximum transparency pursuant to requirement 4 including full disclosure of data sets, decisions made by the system, as well as their unlimited storage does come at the risk of violating other rights. Here, the right to privacy (Art 8 EU Charter) and right to erasure (“right to be forgotten”, Art 17 GDPR) of other users whose personal data is to be made transparent is likely to be violated as well as the developer’s right to keep trade secrets unknown for source code and data sets. Another example could be found in requirement 2 of technical robustness and safety and the societal and environmental well-being of requirement 6. The AI systems’ hunger for electricity and its consequences for the environment depends on the functions such applications should perform and the number of parameters they contain. The more tasks a system should perform and the safer a system is to be designed through

[86] EU AI Guidelines, 20.

regular backups, mirroring datasets and decisions made, constant monitoring and possibilities to rejection, the more energy it is deemed to consume. Such concurrences can be identified with the interplay of numerous requirements from the guidelines and are imposing on consideration with an increasing weighing of the rights and principles in question.

2) *High operational costs and the risk of overburdening AI developers*

Secondly, with this culmination of making developers and operators of AI systems responsible for everything and everyone but leaving them with the weight of uncertain trade-offs, the guidelines risks overburdening the very actors that are supposed to bring forward market-ready AI. Moreover, as Whittaker points out, the guidelines do not discuss how the additional costs of Trustworthy AI systems would be allocated.[87] These costs are likely to be substantial as we can see from the rise of operational costs for services falling under the GDPR application.[88] With the greatly expanded scope of protection and the increased requirements for risk assessments, it must be taken for granted that the costs for compliance with the requirements will surpass the costs for GDPR by far. That is striking because most non-public AI applications are designed from and operated for companies that are required to offer their services for competitive prices. Since the guideline's requirements do not necessarily add direct value to their service, AI providers will have a hard time allocating the additional costs to their services streams. This burden of shouldering higher costs than non-compliant domestic or AI competitors from other legislations is likely to result in a substantial drawback in developing and implementing market-ready AI applications.

As practical consequences, developers could either cherry-pick or even discard the requirements as academic and too costly and refrain from embodying them in their AI projects, or comply with them but face a competitive disadvantage. In the first case, the developers would not suffer from higher administrative costs than their competitors, enabling them to participate in the global race for creating bringing market-ready AI applications more evenly. The probable AI systems brought forward would then lack compliance with the guidelines and shape non-compliant market standards so that the guidelines run dry. AI applications which, in contrary, embody the guidelines could fail to prevail in this market

[87] *Whittaker*, J.I.B.L.R. 2019, 300.

[88] Ernst & Young estimated in 2018 that the costs for complying with the GDPR for Fortune 500 companies accumulated to USD 8 Billion, see https://www.bloomberg.com/news/articles/2018-03-22/it-ll-cost-billions-for-companies-to-comply-with-europe-s-new-data-law.

because of their increased costs or impaired functionality. This risk must be seen in the context of how setting digital standards is a global race nowadays. Falling behind at this early stage in the development of an aspiring technology will likely result in leaving the field for other actors that set their own standards. Once these actors and their applications have reached the critical mass market share and penetration, such applications scale globally and reach monopoly-like market share, putting competitors in the situation of constantly running after the dominant company. If European AI fails that in that competition setting standards globally, trustworthy AI made in the EU could end like a paper tiger that lives in the shadow of global predators. It remains to be seen if developers embodying the demanding Guidelines principles into their AI applications will prevail in the current global rat race.

3) *A starting point with an emphasis on transparency and accountability*

On the other hand, the guidelines emphasise principles that are crucial for exploiting the technology's potential while fostering human rights and ecological sustainability. Mindful of the fact that the requirements are not legally binding and remain abstract, their principles could perhaps rather be seen as a starting point for designing responsible AI systems. [89] From that point of view, it seems reasonable pointing out the basis of a state under the rule of law whose institutions and stakeholders are directed by ethical and environmental requirements. Both guidelines urge the same interlocking principles for that reason by addressing AI's autonomy, data quality, process quality, transparency, and accountability.[90]

In such light, it appears to be understandable to, for instance, particularly emphasise the need for creating AI applications whose decisions are no Blackbox but are explainable. Tackling the current black-box architecture is rightfully seen as one of the main obstacles that stand in the way of acquiring stakeholder's trust in delegating tasks and decisions to machines.[91] The same holds true for accountability for AI systems developed and deployed as well as their auditability and reporting of impact, trade-offs and redress. From the author's view, AI systems do not have to be identified as persons that can be held responsible (yet), as truly autonomous machines are still up in the air.[92] What is necessary to build trust in AI applications that are increasingly autonomous in solving certain tasks and formulating

[89] *Salami*, C.T.L.R. 2020, 127; *Whittaker*, J.I.B.L.R. 2019, 300; *Clarke*, C.L. & S.R. 2019, 410, 415.

[90] *Clarke*, C.L. & S.R. 2019, 410, 415 f.

[91] *Yu/Ali*, L.I.M. 2019, 6 f.; *Clarke*, C.L. & S.R. 2019, 423, 428 f.; *Zuckermann*, L.Q.R. 2020, 451 f.

[92] But see the discussion: *Bhargava/Velasquez*, G.J.L. & P.P. 2019, 831 ff.

consequences, instead, is having a comprehensive set of accountability rules in place with which developers and appliers of AI systems can be held responsible for the machines' deeds.[93] Of course, AI accountability can only be enforced where necessary information is to be obtained through the obligation to audits and reports. The focus on transparent decisions and accountability is rightly made and should be become a cornerstone of AI developments.

4) *Laissez-faire, preventive regulation, or identifying risk sectors*

The dominant question for setting the course for AI made in the EU will be the political decision for a liberally regulated laissez-faire playing field, more preventive regulation, or a risk-based solution in-between. Salami argues that the EU requires further regulation and guidance for AI systems. [94] A number of requirements from the EU AI Guidelines are already reflected in existing laws such as the principles of privacy and accountability through the GDPR. In his view, other requirements say the principles of transparency and human oversight do lack legislative safeguards. Sectors that are lacking EU-wide regulation could be menaced by inconsistent national regulation, fragmenting the internal markets of the EU for AI systems.[95] In order to provide more clarity as to what is expected in the application of existing laws to AI, to encourage uniform applicability and regulation of identified gaps, more specific guidance should be brought forward.[96]

As was the case in identifying the legal foundations for AI regulations,[97] in the author's view, it could be useful to differentiate between high risk applications and no high risks.[98] The first mentioned would then be subject to a higher standard of regulation than the latter. As a result, proportionality between the specific AI application's risks and the correlating regulation could be fostered. That approach acknowledges the different risks for stakeholders associated with different sectors of AI applications say in healthcare and transport on one hand and digital entertainment on the other hand. Regulation in high risk sectors should shine a light on training data, data and record-keeping, information to be

[93] In that sense, the current discussion shows some parallels with the historical debate around corporate legal responsibility for violations of rights committed from within corporations: *Bhargava/Velasquez*, G.J.L. & P.P. 2019, 834 f. ,849 f.

[94] *Salami*, C.T.L.R. 2020, 126.

[95] *Salami*, C.T.L.R. 2020, 126.

[96] *Salami*, C.T.L.R. 2020, 126.

[97] See above under: Chapter 2, II. 1).

[98] As has recently been suggested by the European Commission, see EU AI Whitepaper, 16 f., 24.

provided, robustness and accuracy, human oversight, specific requirements for certain AI applications such as those used for purposes of remote biometric identification.[99] For no high risk applications it seems reasonable to help foster innovation without such requirements first and give developers the possibility to voluntarily apply higher standards. In this case, it seems reasonable to set up certification and accreditation standards so that optional good industry practice can receive attention and can act as a competitive edge. In both situations the AI developing industry would enjoy the clarity as to which development and deployment can be undertaken and under which requirements.

On the other hand, Salami argues, that such regulatory approach may result in clustering regulation with some AI applications escaping the regulative regime.[100] And clearly, the idea to identify high and no high risk sectors for a technology whose potential is yet unknown leaves plenty of room for speculation and wrong prognosis which could, ultimately, cause adverse effects for stakeholders of AI applications. In his view, a better approach would be to create a general legal framework applicable to all AI applications.[101] From such framework, however, only the relevant provisions suitable for the specific application should find application. In my view, that appears to be a river crossing without getting wet. On the one side, the one size fits all approach requires a regulatory framework applicable to all AI systems regardless of their nature and profile so that the specific risks are no requirement for its application. On the other side, Salami suggests interpreting the general framework as a set from which only certain tools are to be applied by the industry depending on the specific circumstances of the AI application. While such approach offers more flexibility, it holds practical difficulties namely the questions who is going to apply and properly monitor the rules so that they meet the specific AI systems profile.

Also, factually, the solution does not appear to be so different to the risk-based approach because its application and monitoring will work with similar categories of risk as to which specific AI applications will have to meet different regulatory requirements. In the author's view, Salami's approach appears to be a botched job, demanding more regulation and flexibility while resulting in less clarity and more administrative burden. A strong laissez-faire approach would not seem preferable either as it would likely result in a lack of

[99] EU AI Whitepaper, 18.

[100] *Salami*, C.T.L.R. 2020, 127.

[101] *Salami*, C.T.L.R. 2020, 127.

attention for AI’s harmful potential. Therefore, endorsing the risk-based approach appears to be the best strategy in addressing AI’s inherent risks while leaving developers enough room for invention. In that context, Clarke convincingly argues that policymakers and organisations should design such risk assessments so that all stakeholders are taken into account.[102] In other words, they should be broader than usually undertaken within organisations. Additionally, such assessments should be conducted from the perspective of each stakeholder group.[103]

5) *Summary*

Perhaps it is yet too early to expect an approach to responsible AI to be equally comprehensive and contingent. With the regulative attempts illuminated, one cannot help but notice what giant efforts it will require to create suitable regulation for AI systems and their development. The matter is complex and will have to take into account all sorts of political, financial and legal measures. Still, the UK and EU AI Guidelines underline the significance of AI’s consequences for stakeholders and demand attention. Because they are neither legally binding nor exhaustive or fully developed, the guidelines and their principles should be seen as a comprehensive set that proposes a starting point for designing responsible AI systems.[104] Their focus is to bring forward aspects of an AI system architecture that can be “ethical by design”. While the fluid and sometimes even contradicting nature of the principles as well as passing on of trade-offs to the developers are reasonable objections to the guidelines’ application, they address the interplay of ethical questions concerned with AI usage. With it, developers and operators are given a starting manual to assist them in assessing their AI projects and address how stakeholders are being affected with them; an opportunity to improve AI applications that did not exist before. Moreover, the guidelines are laying a foundation for future legislation, technical norms, and private schemes addressing responsible usage of AI. The first example for such regulation of AI has been found in Art 13(2)(f) and 22 GDPR.

[102] *Clarke*, C.L. & S.R. 2019, 410, 413.

[103] *Clarke*, C.L. & S.R. 2019, 410, 413.

[104] *Salami*, C.T.L.R. 2020, 127; *Whittaker*, J.I.B.L.R. 2019, 300.

II. Premises crucial for the development of responsible AI

Mindful of the fact that AI is still bound by the "disposition of its organs"[105] and against the background of the discussion above, the following premises appear to be central in the author's view for fostering responsible but competitive AI. These balanced dispositions should guide both policymakers and the industry.

1. Identifying risk sectors for AI usage to encourage sector-specific regulation.
2. Refraining from overburdening AI developers with extensive preventive regulation in low-risk sectors.
3. Promoting specific requirements that are essential for responsible AI developments, specifically explainability and accountability for decisions made.
4. Integrating the industry in setting up standards for responsible AI early on.

The policymakers must be wary not to become overzealous in their rightful attempt to direct the development of that promising technology while remaining vigilant in safeguarding the rules and expectations of humane societies. The industry should live up to its role as a stakeholder of society and actively participate in the development of an AI which is not just economically promising, but which is also responsible, transparent, and accountable.

Conclusions

Today, AI is being associated with unlimited possibilities and is expected to transform society as a whole. While AI is still in its infancy, it bears substantial risks to its stakeholders. These risks are being addressed by the European attempts for directing and regulating AI. Thereby, the EU AI Guidelines and the UK AI Guide do emphasise the importance of the ethical principles of democratic societies under the rule of law. It seems yet too early to expect such guidance to be equally comprehensive and contingent. Still, the guidelines provide a highly expected starting point for a critical approach to AI, indicating the motives and direction of future legislation. The GDPR provisions on the effects of algorithmic decision-making on fundamental rights constitute the first step in that direction. In the author's opinion, the regulators should not overburden AI developers with extensive preventive regulation but identify risk sectors for AI developments first. However, certain principles of responsibility, namely, explainability and accountability for decisions made by

[105] *Descartes*, Discourse 1637, 44.

the AI, should become mandatory requirements for AI systems. Mindful of the fact that AI systems which are ethical by design are still up in the air, the European guidelines and regulation should not be seen as the big leap but as a first step towards shaping a framework for responsible AI systems.

Bibliography

Arruda, A.J.T.A. 2017	Arruda, Andrew, An Ethical Obligation to Use Artificial Intelligence: An Examination of the Use of Artificial Intelligence in Law and the Model Rules of Professional Responsibility, American Journal of Trial Advocacy 2017, 40, 443-458
Bhargava/Velasquez, G.J.L. & P.P. 2019	Bhargava, Vikram R/Velasquez, Manuel, Is Corporate Responsibility Relevant to Artificial Intelligence Responsibility, Georgetown Journal of Law & Public Policy, 2019(17), 829-852
Bootle, 2019	Bootle, Roger, The AI Economy, Work, Wealth and Welfare in the Robot Age, 2019
Campolo/Crawford, E.S.T.S. 2020	Alexander Campolo/Kate Crawford, Enchanted Determinism/ Power without Responsibility in Artificial Intelligence, Engaging Science, Technology, and Society 6 (2020), 1-19
Clarke, C.L. & S.R. 2019, 410	Clarke, Roger, Principles and business processes for responsible AI, Computer Law & Security Review 2019 (35), 410-422
Clarke, C.L. & S.R. 2019, 423	Why the world wants controls over Artificial Intelligence, Computer Law & Security Review 2019 (35), 423-433
Coeckelbergh, S.E.E. 2019	Coeckelbergh, Mark, Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Science and Engineering Ethics (2019)
Descartes, Discourse 1637	Descartes, René, Discourse on the Method, 1637
EU AI Guidelines	Independent High-level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, April 2019
EU AI Whitepaper	WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust, 19.2.2020, COM(2020) 65 final, to be found at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
Flett/Wilson, C.T.L.R. 2017	Flett, Emma/Wilson, Jennifer, Artificial intelligence: is Johnny 5 alive? Key bits and bytes from the UK's robotics and artificial intelligence inquiry, Computer and Telecommunications Law Review, 2017, 23(3), 72-74
House of Commons, AI Report 2016	House of Commons, Science and Technology Committee, Robotics and Artificial Intelligence, Fifth Report of Session 2016-17, to be found at: https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf
House of Lords, AI report 2017	House of Lords, Select Committee on Artificial Intelligence, Report of Session 2017-19, AI in the UK: ready, willing and able?
Kemp, Comms.L. 2019	Kemp, Roger, Regulating the safety of autonomous vehicles using artificial intelligence, Communications Law 2019, 24(1), 24-33
McCarthy et. al., 1955	McCarthy, John/Minsky, Marvin L./Rochester, Nathaniel/Shannon, Claude E., A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, reprinted in AI Magazine 2006, 27(4)

McCarthy, 2007	McCarthy, John, What is artificial intelligence? Department of Computer Science, Stanford University, 2007, at http://www-formal.stanford.edu/jmc/whatisai/node1.html
McGregor et. al., I.C.L.Q. 2019	McGregor, Lorna/ Murray, Daragh/Ng, Vivian, International human rights law as a framework for algorithmic accountability, International & Comparative Law Quarterly, 2019, 68(2), 309-343
Rowe, J.P.I.L. 2018	Rowe, Kurt, The rise of the machines: a new risk for claims?, Journal of Personal Injury Litigation 2018, 4, 302-307
Salami, C.T.L.R. 2020	Salami, Emmanuel, Europe's readiness for the AI takeover: some salient points and comments from the European Commission's white paper on AI, Computer and Telecommunications Law Review, 2020, 26(5), 126-127
Turing, Mind 1950	Turing, Alan M., Computing Machinery and Intelligence, Mind, 1950 (49), 433-460, to be found at: https://www.cs.mcgill.ca/~dprecup/courses/AI/Materials/turing1950.pdf
UK AI Guide	Office of Artificial Intelligence, A guide to using artificial intelligence in the public sector, June 2019
UN CESCR, Comment No. 3	Committee on Economic, Social and Cultural Rights, General Comment No. 3: The Nature of States Parties’ Obligations (Art. 2, Para. 1, of the Covenant), 1990, UN Doc CESCR/E/1991/23
UN HCR, Ruggie Principles	Human Rights Council, Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie, 2011, UN Doc A/HRC/17/31
UN HRC, Comment No. 31	UN Human Rights Committee, General Comment No. 31, The Nature of the Legal Obligation Imposed on States Parties to the Covenant, 2004, UN Doc CCPR/C/21/Rev.1/Add. 13
Washington Post, June 11, 2020	Jay Greene, Microsoft won’t sell police its facial-recognition technology, following similar moves by Amazon and IBM, https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/
Whittaker, J.I.B.L.R. 2019	Whittaker, Andrew, Artificial Intelligence – the new EU guidelines, Journal of International Banking Law and Regulation 2019, 34(9), 295-300
Yu/Ali, L.I.M. 2019	Yu, Ronald/Ali, Gabriele S., What’s inside the black box? AI challenges for lawyers and researchers, Legal Information Management 2019, 19(1), 2-13
Zech, GRUR Int. 2019	Zech, Herbert, Artificial Intelligence: Impact of Current Developments in IT on Intellectual Property, GRUR Int. 2019, 1145-1147
Zittrain, 2019	Zittrain, Jonathan, Intellectual Debt: With Great Power Comes Great Ignorance, Berkman Klein Center for Internet & Society at Harvard University 2019, at https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c
Zuckermann, L.Q.R. 2020	Zuckermann, Adrian, Artificial intelligence - implications for the legal profession, adversarial process and rule of law, Law Quarterly Review, 2020, 136(Jul), 427-453