Reports

International

AI Regulation and Governance on a Global Scale: An Overview of International, Regional and National Instruments

*Mark D. Cole**

I. Introduction

'Like electricity in the past, artificial intelligence (AI) is transforming our world'.[1] As has been the case with any new technology, the existing and future applications of AI systems are raising concerns, in the public, with legislators, in the industry itself and with academics. To address these concerns and the challenges posed by the usage of AI systems, the question is no longer *whether* AI will be regulated but rather *how*. In that line, one could observe numerous national initiatives concerning AI which typically started out with AI expert advisory groups asked to identify the risks associated with the development and deployment of AI systems and provide recommendations to mitigate these risks. This resulted in the emergence of national AI strategies on a broader scale from 2017 onwards.[2]

Moreover, early governance approaches evolved in the form of policy frameworks focussing on the encouragement of innovation and investment in the field.[3] In addition, in international organisations recommendations were developed which aimed at ensuring a use of such systems which people can trust. Notably, the 'OECD Recommendation on Artificial Intelligence' of May 2019 which the OECD (Organisation for Economic Co-operation and Development) considers as a 'global reference point for trustworthy AI'[4] marks a first step towards regulation in the form of legal standards and international agreements. The core value-based principles to lead the trustworthy deployment, development and use of AI identified in the Recommendation have indeed become an influential reference point in subsequent policy frameworks and emerging AI-specific regulation. In 2023, the 'OECD.AI Policy Observatory' enlisted over 50 national strategic and government-wide initiatives on trustworthy AI.[5] In sum, over 930 related policy initiatives across 70 jurisdictions had been reported to the OECD.AI policy hub by May 2023.

Regulation is, however, still in its infancy. Differing views of how to regulate AI exist, but there seems to be consensus that ethic codes alone are not sufficient to manage the risks involved. Questions are raised as to which would be the right level or context in which the topic of regulation should be addressed.

States are exploring regulatory approaches, while soft law in the format of national ethics frameworks and principles[6], professional guidelines[7], private

DOI: 10.21552/aire/2024/1/16

* Mark D. Cole is Professor for Media and Telecommunication Law at the University of Luxembourg and Director for Academic Affairs at the Institute of European Media Law (EMR). He is also founding and associate editor of the European Data Protection Law Review (EDPL) and one of the founding editors of the Journal of AI Law and Regulation (AIRe). Research for this contribution was supported by the project LAIWYERS, funded by the Institute of Advanced Studies (IAS) of the University of Luxembourg. The author would like to thank Dr Sandra Schmitz-Berndt, Research Associate on that project, for assistance in preparing the report. For correspondence: <mark.cole@uni.lu>.

1 European Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Coordinated Plan on Artificial Intelligence' COM(2018) 795 final, Introduction.

2 For a timeline of national AI strategies see OECD, 'The state of implementation of the OECD AI Principles four years on' (OECD Artificial Intelligence Papers No 3, October 2023), 13 <https://rb.gy/tzw5fh>. All Internet links in this report were last accessed 13 March 2024.

3 See, for example, the European Commission's AI4EU programme or the SME 4.0 Excellence Centres in Germany.

4 OECD, 'The state of implementation of the OECD AI Principles four years on' (n 2), 8.

5 Ibid.

6 EU Ethics guidelines on AI (2018); UNESCO Recommendation on the Ethics of AI (2021); OECD Recommendation of the Council on AI (2019); Australia, AI Ethics Framework (2019); Switzerland, Guidelines on AI for the Confederation (2020).

7 For example in the context of using AI in journalism; for another example World Employment Confederation, 'Code of Ethical Principles in the use of artificial intelligence' (2023) <https://wecglobal.org/uploads/2023/04/AI-principles-WEC-AI-Code-of-Conduct-March-2023.pdf>.

standards[8] as well as codes of conduct[9] is emerging globally. In contrast, there is still only a limited amount of hard law approaches. Regulation rather addresses AI applications by sector or domain,[10] without providing for comprehensive regulation at horizontal level. In view of the latter, however, a major development has emerged in the EU, with the EU striving for a legislation that may serve as a blueprint for AI regulation in a similar way in which the GDPR did for the field of data protection.[11] With the 2021 Proposal for an AI Act – the Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts[12] – the European Commission envisaged the adoption of the first comprehensive legal framework regulating AI systems. The Regulation, which is finalised and will shortly be finally voted on, seeks to encourage innovation and deployment of AI while at the same time managing associated risks. The plenary vote in the European Parliament on the AI Act is scheduled for 13 March 2024, three years after the Commission published its Proposal. Once adopted, the AI Act will enter into force 20 days after publication in the Official Journal of the EU but will only become applicable after a transitional period of two years for most parts and for certain elements already after six or twelve months and partly only after three years.

In 2022, 'artificial intelligence took centre stage in policy discussions',[13] when the release of OpenAI's Chat GPT (based on GPT 3.5)[14] exposed the potential of generative AI technologies to the general public. This development meant that the Proposal underwent fundamental changes in the legislative procedure with the aim of responding to these newly apparent risks and which resulted in a separate chapter addressing general purpose AI (GPAI) models. With the concerns raised as regards more powerful GPAI tools, not only the question of the level of regulation becomes pertinent but also the point of time when regulation should ideally become effective. Against this background, this report provides an overview of regulatory approaches at global, supranational/regional and national level. It follows in the first part (II.) a chronological order before taking a separate look at the regional approach of the Council of Europe as the planned Convention has the aim of becoming a potentially global standard which will be binding for signatories (III.). After two examples of national approaches in the United States of America and the United Kingdom (IV.), the most advanced regulatory approach is presented with the AI Act of the European Union (V.). In this regard, the development from policy instruments to hard law efforts is presented and reoccurring and key notions in AI regulation are identified. The contribution shall serve the purpose of being able to compare the different approaches with the goal of identifying commonalities and differences (see also part VI.). To facilitate this, there is an Annex to the contribution with a summary table overview of key points.

II. Approaches on the Global Level

1. The OECD Recommendation of the Council on Artificial Intelligence

The OECD's goal is to create better policies for better lives and to 'shape policies that foster prosperity, equality, opportunity and well-being for all'.[15] In that regard, the OECD for instance had adopted the first global framework for data protection in 1980.[16] Since 2016, the OECD has undertaken empirical and policy activities on AI in support of the policy debate,

8 NIST AI Risk Management Framework (2023); ISO/IEC 23053:2022 Framework for AI Systems Using Machine Learning (2022); CEN/CENELEC standards for AI and related data (forthcoming).

9 G7, 'Hiroshima Process International Code of Conduct for organisations developing advanced AI systems' (2023); Proposal for a US-EU AI Code of Conduct.

10 Legislation has so far inter alia addressed the regulation of AI-related aspects in data protection (eg art 22 GDPR) and media-specific consumer protection law (eg art 28b AVMSD (Audiovisual Media Services Directive 2010/13/EU as amended by Directive (EU) 2018/1808) imposing certain measures on Video-Sharing Platforms in view of their use of algorithmic systems for the presentation of content).

11 See on the potential 'Brussels Effect' below at VI.

12 European Commission, 'Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts' COM(2021) 206 final.

13 OECD, 'The state of implementation of the OECD AI Principles four years on' (n 2), 8.

14 See <https://openai.com/chatgpt> and Dela Cruz, 'ChatGPT Timeline: Evolution and Rise of AI, Impact, Threat, and Opportunities' (28 February 2023) <https://www.techtimes.com/articles/287927/20230228/chatgpt-timeline-evolution-rise-ai-impact-threat-opportunities.htm>.

15 See OECD, 'Who we are' <https://www.oecd.org/about/>.

16 OECD, 'OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data' (1980) OECD/LEGAL/0188.

starting with a Technology Foresight Forum on AI in the same year.

Based on the proposal of an expert group initiated by the Committee on Digital Economy Policy (CDEP), the OECD Council adopted the Recommendation on Artificial Intelligence[17] at its meeting at Ministerial level on 22-23 May 2019. All 38 OECD members[18] as well as eight non-OECD members[19] have adhered to the Recommendation which constitutes the first intergovernmental standard on AI policy.

The Recommendation aims to foster innovation and trust in AI by promoting the responsible stewardship of trustworthy AI while ensuring respect for human rights and democratic values. Complementing existing OECD standards in areas such as privacy, digital security risk management, and responsible business conduct, the Recommendation sets out a framework containing ten principles – divided into (1) five value-based principles for the responsible stewardship of trustworthy AI (section 1) and (2) five recommendations to governments to promote and implement in their policies said responsible stewardship of trustworthy AI (section 2).

In view of a consistent interpretation of the framework, an introduction provides definitions for 'AI system', 'AI system lifecycle', 'AI knowledge', 'AI actors and stakeholders'.[20] For the purpose of the Recommendation, 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'[21]. The Recommendation recognises that the obligations of AI actors, encompassing 'those who play an active role in the AI system lifecycle'[22] do not end with system deployment, but that the AI system lifecycle also encompasses the operation and monitoring phase, that follows the development, verification and validation, and deployment phases.[23]

Section 1 of the Recommendation identifies five value-based principles for responsible stewardship of trustworthy AI and calls on all AI actors to promote and implement them. In addition to these complementary principles, which will further be addressed below, Section 2 of the Recommendation encompasses five recommendations to policy-makers relating to national policies and international cooperation for trustworthy AI. The adherents to the recommendation are advised to (1) implement in their national policies public investment and encourage private investment in AI research and development; (2) foster a digital ecosystem for AI; (3) shape an enabling policy environment for AI that also encourages innovation and competition; (4) build human capacity and (5) prepare for labour market transformation.

As regards international co-operation, the adherents should actively co-operate to advance the Section 1 principles, share AI knowledge and promote the development of global technical standards for interoperable and trustworthy AI. Furthermore, the development of metrics to measure AI research, development and deployment, and for building an evidence base to assess progress in its implementation is recommended.

In consideration that the Recommendation mentions the notion of 'trustworthy AI' or 'trustworthiness of AI systems' 22 times, the principles enshrined in the first section introduce the concept of 'responsible stewardship for trustworthy AI'. All AI actors are called upon to promote and implement the OECD AI principles, which encompass (1) inclusive growth, sustainable development and well-being; (2) human-centred values and fairness; (3) transparency and explainability; (4) robustness, security and safety; and (5) accountability for the proper functioning of AI systems and respect of the aforementioned principles.

Four years after the adoption of the Recommendation, the OECD infers that its principles are serving as a global reference point for trustworthy AI and are being translated into 'concrete, operational initiatives'.[24] Following an assessment of the state of

17 OECD, 'Recommendation of the Council on Artificial Intelligence' (2019) OECD/LEGAL/0449.

18 Australia, Austria, Belgium, Canada, Chile, Colombia (adherence still as non-member), Costa Rica (adherence still as non-member), Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Republic of Turkey, United Kingdom, and the United States.

19 Argentina, Brazil, Egypt, Malta, Peru, Romania, Singapore, and Ukraine.

20 OECD, 'Recommendation of the Council on Artificial Intelligence' (n 17), para I.

21 Ibid.

22 Ibid.

23 Ibid.

24 OECD, 'The state of implementation of the OECD AI Principles four years on' (n 2), 8.

implementation of the OECD AI principles four years on[25], the OECD took the opportunity to reassess the definition of 'AI system' in light of the emergence of general purpose AI models. On 8 November 2023, a revised definition of 'AI system' was adopted in order to ensure that the Recommendation 'continues to be technically accurate and reflect technological developments, including with respect to generative AI'.[26] According to the updated Recommendation an 'AI system' is defined as 'a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment'.[27]

The updated definition includes edits aimed at clarifying the objectives of an AI system (which may be explicit or implicit); underscoring the role of input which may be provided by humans or machines; clarifying that the Recommendation applies to generative AI systems, which produce 'content'; as well as reflecting the fact that some AI systems can continue to evolve after their design and deployment.[28] Furthermore the word 'real' is replaced by 'physical' in order to align the definition with other international processes such as the definition of an AI system proposed by the Council of Europe in its Draft for a Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (see below at III.).[29]

Besides revising the Recommendation, the OECD is continuing its work on AI policy. In February 2022, a Framework for Classifying AI systems[30] was published to guide stakeholders including legislators in assessing the opportunities and risks of different types of AI systems. The Framework acknowledges that AI systems depending on their context raise different challenges and links AI systems characteristics with the OECD AI principles.[31]

2. The G20 AI Principles

At the Osaka Summit in June 2019, the G20[32] expressed their commitment to a 'human-centred' approach to AI, which is guided by the 'G20 AI Principles'[33]. The G20 AI principles draw upon the OECD Recommendation and recite (1) the principles for responsible stewardship of trustworthy AI and (2) the recommendations for national policies and international cooperation for trustworthy AI included in section 1 and 2 of the OECD Recommendation.[34] Departing from the OECD Recommendation, the G20 AI Principles abstain from restating the definitions agreed upon in the OECD Recommendation and also refrain from providing separate definitions for example for an 'AI system'.

The commitment to the G20 AI Principles has recently been reaffirmed under the heading 'Harnessing Artificial Intelligence Responsibly for Good and for All' during the 18th G20 Heads of State and Government Summit in September 2023.[35]

3. The UNESCO Recommendation on the Ethics of Artificial Intelligence

The first truly global – yet non-binding – normative framework emerged on 23 November 2021 with the 'UNESCO Recommendation on the Ethics of Artifi-

25 Ibid.

26 OECD, 'Recommendation of the Council on Artificial Intelligence' (n 17), Background Information.

27 Ibid, as amended on 8 November 2023.

28 Ibid.

29 Council of Europe, CAI, 'Revised Zero Draft [Convention] on Artificial Intelligence, Human Rights, Democracy and the Rule of Law' (6 January 2023), CAI(2023)01.

30 OECD, 'The OECD Framework for the Classification of AI systems' (OECD Digital Economy Papers No 323, 2022) <https://rb.gy/ubc0vp>.

31 The categories people and planet, economic context, data and input, AI model, task and output, are identified as the key dimensions along which AI systems and applications have to be classified. Each of these dimensions has distinct properties and attributes, which are relevant for a nuanced policy assessment (ibid, 16 et seq).

32 G20 (Group of Twenty) is an informal gathering of 19 States representing the largest part of the global GDP and contributing to around three quarters of world trade plus two regional organisations (EU and AU). In its self-description it 'is the main forum for international economic cooperation. It plays an important role in defining and strengthening global architecture and governance on all major international economic issues', see <https://www.g20.org/en/about-the-g20>.

33 G20, 'G20 Ministerial Statement on Trade and Digital Economy' (June 2019) <https://www.mofa.go.jp/files/000486596.pdf>.

34 G20, 'G20 AI Principles' Annex to the G20 Ministerial Statement on Trade and Digital Economy (June 2019) <https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf>.

35 G20, 'G20 New Dehli Leaders' Declaration' (10 September 2023), para 61 <https://www.mea.gov.in/Images/CPV/G20-New-Delhi-Leaders-Declaration.pdf>.

cial Intelligence'[36]. The UNESCO (United Nations Educational, Scientific and Cultural Organisation) has 194 Member States and 12 Associate Members, while at the time of adoption of the Recommendation it had 193 members.[37] The Recommendation addresses ethical issues of AI technologies to the extent that they are within UNESCO's mandate. The general conference standard-setting instrument is aimed at States[38] and indirectly stakeholders[39] to guide the construction of the necessary legal infrastructure to ensure a so-called healthy development of AI.[40] An Ad Hoc Expert Group was tasked in March 2020 to draft a global ethical framework that highlights the advantages of AI while reducing the risk that AI technologies entail.[41]

Other than the OECD Recommendation, the UNESCO instrument refrains from providing a definition of AI, arguing that such a definition would need to be updated over time in line with technological developments.[42] Instead, the Recommendation emphasises that it addresses 'those features of AI systems that are of central ethical relevance' and in that regard approaches AI systems as 'systems which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control'. [43]

In the following, three elements are enlisted which are attributed a 'central place in this approach', which means that in view of these elements there are ethical issues identified. The first element enlists features of AI systems that relate to how an AI system works, referring inter alia to machine learning or machine reasoning models with varying degrees of autonomy to perform cognitive tasks. The second element addresses the AI life cycle and the actors involved in the life cycle, while the third element addresses new types of ethical issues raised by AI including their impact on decision-making and various environments.

All actors in the AI system life cycle are called to respect the following complementary and interrelated (1) values and (2) principles.[44] The values are intended to inspire 'desirable behaviour' and represent the foundation of principles, which 'unpack the values underlying them more concretely so that the values can be more easily operationalised' in concrete actions.

The values to be considered are (1) respect, protection and promotion of human rights and fundamental freedoms and human dignity;[45] (2) environment and ecosystem flourishing;[46] (3) ensuring diversity and inclusiveness;[47] as well as (4) living in peaceful, just and interconnected societies[48]. The principles to be respected consist of (1) proportionality and 'do no harm';[49] (2) safety and security;[50] (3) fairness and non-discrimination;[51] (4) sustainability;[52] (5) right to privacy and data protection;[53] (6) human oversight and determination;[54] (7) transparency and explainability;[55] (8) responsibility and accountability;[56] (9) awareness and literacy;[57] as well as (10) multi-stakeholder and adaptive governance and collaboration[58].

36 UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (2022) SHS/BIO/PI/2021/1.

37 After the US rejoined on 10 July 2023, UNESCO has 194 members today, see <https://www.unesco.org/en/countries>.

38 UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (n 36), para 4.

39 Cf ibid, paras 4 and 48.

40 UNESCO, 'UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence' (25 November 2021) <https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence>.

41 UN, '193 countries adopt first-ever global agreement on the Ethics of Artificial Intelligence' (25 November 2021) <https://news.un.org/en/story/2021/11/1106612>.

42 UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (n 36), para 2.

43 Ibid.

44 Para 10.

45 Paras 13 et seq.

46 Paras 17 et seq.

47 Paras 19 et seq.

48 Paras 22 et seq.

49 Paras 25 et seq.

50 Para 27.

51 Paras 28 et seq.

52 Para 31.

53 Paras 32 et seq.

54 Paras 35 et seq.

55 Paras 37 et seq.

56 Paras 42 et seq.

57 Paras 44 et seq.

58 Paras 46 et seq.

The Recommendation also identifies as a third overarching element concrete areas of policy action that operationalise the aforementioned values and principles. These policy areas of practical realisation, reflecting the main fields of activity of UNESCO, are (1) ethical impact assessment;[59] (2) ethical governance and stewardship;[60] (3) data policy;[61] (4) development and international cooperation;[62] (5) environment and ecosystem;[63] (6) gender;[64] (7) culture;[65] (8) education and research;[66] (9) communication and information;[67] (10) economy and labour;[68] and (11) health and social well-being.[69] Notably, the Recommendation establishes in the policy section practical methodologies to support the effective implementation of the Recommendation.

Upon request of Member States, UNESCO has developed a Readiness Assessment Methodology[70] as well as a complementary Ethical Impact Assessment tool[71] to assist states in building their capacities.

As regards the protection of individuals, the protection of personal data, the banning of social scoring and mass surveillance have been identified as key achievements of the UNESCO AI Recommendation.[72] Furthermore, the Recommendation bans the replacement of humans by AI system when it comes to life and death decisions.[73] Other than the OECD Recommendation, the UNESCO Recommendation puts great emphasis on environmental and ecosystem protection,[74] thereby recognising that AI may become a prominent tool in the fight against climate change and other environmental issues.[75] It further calls for a peaceful use of AI systems.

4. The G7 Hiroshima AI Process Comprehensive Policy Framework

The 'Hiroshima AI Process Comprehensive Policy Framework' constitutes the output of the so called 'Hiroshima AI Process' which was launched in May 2023 under Japan's G7[76] Presidency and focussed on the opportunities and challenges of generative AI technologies. As such, it is a direct response to a generative AI taking 'centre stage in the public, academic, and political discussions' surrounding this technology.[77] The Policy Framework is promoted by the G7 Leaders as the 'first successful international framework comprising of guiding principles and a code of conduct to address the impact of advanced AI systems on our societies and economies'.[78]

The Hiroshima AI Process Comprehensive Policy Framework consists of four complimentary pillars: (1) the OECD's Report towards a G7 Common Understanding on Generative AI (the 'OECD Report'[79]); (2) the Hiroshima Process International Guiding Principles for all AI actors and for organisations develop-

59 Paras 50 et seq.

60 Paras 54 et seq.

61 Paras 71 et seq.

62 Paras 78 et seq.

63 Paras 84 et seq.

64 Paras 87 et seq.

65 Paras 94 et seq.

66 Paras 101 et seq.

67 Paras 112 et seq.

68 Paras 116 et seq.

69 Paras 121 et seq.

70 UNESCO, 'Readiness assessment methodology: a tool of the Recommendation on the Ethics of Artificial Intelligence' (2023) SHS/REI/BIO/REC-AIETHICS-TOOL/2023.

71 UNESCO, 'Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence' (2023) SHS/REI/BIO/REC-AIETHICS-TOOL-EIA/2023.

72 See Marc Rotenberg, 'Human Rights Alignment: The Challenge Ahead for AI Lawmakers' in Hannes Werthner et al (eds), *Introduction to Digital Humanism* (Springer 2023), 611, 616.

73 UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (n 36), para 36.

74 Cf for instance ibid, paras 17 et seq and 86. In fact, the Recommendation utilises the notion 'environment' 55 times.

75 See Rotenberg (n 72); Josh Cowls et al, 'The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations' (2023) 38 AI & Society, 283–307.

76 The G7 is composed of Canada, United States of America, United Kingdom, Germany, France, Italy and Japan as an informal cooperation forum of the most important industrial powers of the Western world at the time of its foundation; these contribute to around 45 percent of the global GDP. In view of the observer status that the European Commission has, the Hiroshima AI Process also included the EU.

77 OECD, 'G7 Hiroshima Process on Generative Artificial Intelligence (AI) towards a G7 Common Understanding on Generative AI' (7 September 2023), 7 <https://read.oecd-ilibrary.org/science-and-technology/g7-hiroshima-process-on-generative-artificial-intelligence-ai_bf3c0c60-en#page1>.

78 G7, 'G7 Leaders' Statement on the Hiroshima AI Process' (30 October 2023) <https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document01_en.pdf>.

79 OECD, 'G7 Hiroshima Process on Generative Artificial Intelligence (AI) towards a G7 Common Understanding on Generative AI' (7 September 2023) <https://read.oecd-ilibrary.org/science-and-technology/g7-hiroshima-process-on-generative-artificial-intelligence-ai_bf3c0c60-en#page1>.

ing advanced AI systems (the 'Hiroshima Guiding Principles'[80]); (3) the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems[81]; and (4) project-based cooperation in support of the development of responsible AI tools and best practices.[82]

The OECD Report (pillar 1) provides an analysis of priority risks, challenges and opportunities of generative AI and supported the discussions of the G7 Hiroshima AI Process, in particular the Ministers' Meeting on generative AI in September 2023.[83] Furthermore, the Report serves to introduce a common understanding of 'generative AI', namely generative AI as 'a form of AI model specifically intended to produce new digital material as an output (including text, images, audio, video, software code), including when such AI models are used in applications on massive amounts of data; they work by predicting words, pixels, waveforms, data points, etc. that would resemble the models' training data, often in response to prompts'.[84] While the OECD Report refers to generative AI, the further Hiroshima AI Process documents replace this notion with the term 'advanced AI systems'.[85]

The non-exhaustive list of principles in the Hiroshima Guiding Principles (pillar 2), similar to the G20 AI principles, build on the existing OECD AI Principles; however, in contrast to the G20 AI Principles, the Hiroshima Principles do not merely quote the OECD AI Principles, but adapt these principles in consideration of the aforementioned advanced AI systems. The eleven principles enlisted include obligations (1) to conduct risk assessment and appropriate risk management; (2) to identify and mitigate vulnerabilities; (3) to report of advanced AI systems' capabilities, limitations and domains of use; (4) to share information and report incidents; (5) to develop, implement and disclose a risk-based governance approach; (6) to invest in and implement security; (7) to develop and deploy mechanism to enable users to identify AI-generated content; and (11) to implement measures to protect personal data and intellectual property rights. In addition, the G7 members should (8) prioritise research to mitigate AI risks as well as to (9) prioritise the development of advanced AI systems for the common good (including Sustainable Development Goals). Furthermore, the (10) advancement of international technical standards is encouraged. The G7 Digital and Tech Ministers add as a twelfth principle the promotion and contribution to trustworthy and responsible use of advanced AI systems to the non-exhaustive list.[86]

By way of a voluntary guidance, the Hiroshima Code of Conduct (pillar 3) provides a list of actions and recommendations for organisations developing advanced AI systems. This 'living document' is meant to help seize the benefits and address the risks of advanced AI systems. Finally, in relation to project-based cooperation (pillar 4), the G7 welcome existing coordinated efforts[87] to advance trust in generative AI and express their intent to promote and deepen collaboration.

5. The AI Safety Summit Bletchley Declaration

In November 2023, the UK hosted and chaired the inaugural global 'AI Safety Summit' bringing together leading 'AI nations', technology companies, researchers and civil society groups to identify next steps for the development of so called 'frontier AI'.[88]

For the purpose of the Summit, 'frontier AI' was defined as 'highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most ad-

80 G7, 'Hiroshima Process International Guiding Principles for all AI actors and for organisations developing advanced AI systems' (2023) <https://www.mofa.go.jp/files/100573471.pdf>.

81 G7, 'Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems' (2023) <https://www.mofa.go.jp/files/100573473.pdf>.

82 G7, 'Hiroshima AI Process G7 Digital & Tech Ministers' Statement' (1 December 2023) <https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document02_en.pdf>.

83 OECD, 'G7 Hiroshima Process on Generative Artificial Intelligence (AI) towards a G7 Common Understanding on Generative AI' (7 September 2023) p. 2.

84 OECD, 'G7 Hiroshima Process on Generative Artificial Intelligence (AI) towards a G7 Common Understanding on Generative AI' (7 September 2023), 6 based on the understanding of generative AI derived from OECD, 'AI language models: Technological, socio-economic and policy considerations' (OECD Digital Economy Papers No 352, 2023).

85 G7, 'Hiroshima Process International Guiding Principles' (n 80), 1.

86 G7, 'Hiroshima AI Process G7 Digital & Tech Ministers' Statement' (1 December 2023) <https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document02_en.pdf>.

87 Such as the Global Challenge to Build Trust in the Age of Generative AI <https://globalchallenge.ai/>.

88 See UK Government, 'Chair's Summary of the AI Safety Summit 2023, Bletchley Park' (2 November 2023) <https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park>.

vanced models' which include large language models such as those underlying for instance OpenAI's GPT-4.[89] Based on a discussion paper on the capabilities of, and risks arising from, frontier AI[90], the participants exchanged views on the most significant risks and opportunities of frontier AI. Recognising that no single part of society can address the impacts of frontier AI alone, the states[91] and the EU attending the AI Safety Summit, adopted what it referred to as the first international declaration on AI, the 'Bletchley Declaration'[92].

The legally non-binding policy paper stresses 'that AI should be designed, developed, deployed, and used in a manner that is safe, in such a way as to be human-centric, trustworthy and responsible' and emphasises the need 'for the safe development of AI and for the transformative opportunities of AI to be used for good and for all, in an inclusive manner in our countries and globally'.

In that regard, the Declaration sets out an Agenda for addressing 'frontier AI' focussing on (1) identifying AI safety risks of shared concern, building a shared scientific and evidence-based understanding of these risks, and sustaining that understanding as capabilities continue to increase, in the context of a wider global approach to understanding the impact of AI in our societies; and (2) building respective risk-based policies across signatory countries to ensure safety in light of such risks, collaborating as appropriate while recognising that approaches in states may differ based on national circumstances and applicable legal frameworks. This includes, alongside increased transparency by private actors developing frontier AI capabilities, appropriate evaluation metrics, tools for safety testing, and developing relevant public sector capability and scientific research. The AI Summit participants agreed to convene every six months to continue the global conversation on frontier AI.

III. The Council of Europe Draft for a Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law

The Council of Europe considers the area of artificial intelligence as a 'cross-cutting priority'[93] and is pursuing several initiatives to ensure that the application of AI is based on human rights, the rule of law and democracy in a bid to demonstrate 'its ability to pioneer new standards'[94] also in the field of AI. From 2019 to 2021, the Council of Europe's Ad Hoc Committee on Artificial Intelligence (CAHAI)[95] engaged in preparatory work for a possible legal framework to ensure that AI is used to promote and protect the standards of the Council of Europe as an international human rights organisation.[96]

During this period, a Resolution on the Need for Democratic Governance of Artificial Intelligence[97] was adopted by the Parliament Assembly of the Council in 2020, which inter alia called on the Council of Europe member States and other observer States participating in CAHAI 'to work together towards a legally binding instrument'.[98] The Assembly agreed that such an instrument should inter alia 'guarantee that

89 UK Department for Science, Innovation and Technology, 'Capabilities and Risks from Frontier AI' (October 2023), 4 <https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf>.

90 Ibid.

91 Australia, Brazil, Canada, Chile, China, France, Germany, India, Indonesia, Ireland, Israel, Italy, Japan, Kenya, Kingdom of Saudi Arabia, Netherlands, Nigeria, Philippines, Republic of Korea, Rwanda, Singapore, Spain, Switzerland, Turkey, Ukraine, United Arab Emirates, UK of Great Britain and Northern Ireland as well as the United States of America.

92 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023' (1 November 2023) <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>.

93 Council of Europe, 'The Council of Europe & artificial intelligence' (March 2023), 3 <https://rm.coe.int/brochure-artificial-intelligence-en-march-2023-print/1680aab8e6>.

94 Council of Europe, 'The Council of Europe & artificial intelligence' (Foreword by Marija Pejčinović Burić, March 2023) <https://rm.coe.int/brochure-artificial-intelligence-en-march-2023-print/1680aab8e6>.

95 The CAHAI was formally established at the 1353rd meeting of the Committee of Ministers on 11 September 2019, CM/Del/Dec(2019)1353/1.5-app <https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016809737a1>.

96 See Council of Europe, 'Towards regulation of AI systems, Compilation of contributions' (CoE Study DGI (2020)16, December 2020) <https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a>; Council of Europe, CAHAI, 'Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law' (3 December 2021) CAHAI(2021)09rev.

97 Council of Europe, Parliamentary Assembly, 'The Need for democratic governance of artificial intelligence' (22 October 2020) Resolution 2341(2020).

98 Ibid, para 13.

AI-based technologies are designed, developed and operated in full compliance with, and in support of, the Council of Europe's standards on human rights, democracy and the rule of law'[99] and 'provide for the respect of key ethical principles and concepts'[100].

From 2022 onwards, the Committee on Artificial Intelligence (CAI) built upon the work of its predecessor CAHAI, and published the revised zero Draft for a 'Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law'[101] in January 2023. Work continued on the 'zero Draft' and, on 23 December 2023, the Draft Framework Convention[102] was published: it contains the outcomes of the 2nd reading of the text in the Committee and shall serve as the basis for the final reading.

By opting for the legal instrument of a Framework Convention rather than comprehensive regulation in the form of a convention creating rights and obligations, the instrument sets out broad principles, core values and areas for action, seeking to put existing standards on human rights, democracy and the rule of law in the context of AI. Although the instrument is intended to be the first legally binding public international law treaty on AI, as a Framework Convention, significant discretion would be left to signatory states as to how the principles and values would be implemented in practice by the signatories.[103]

As a further principles-based framework, the Draft Framework Convention sets out (1) general obligations, (2) principles related to activities within the lifecycle of AI systems, (3) remedies, (4) obligations regarding the assessment and mitigation of risks as well as (5) guidance to ensure oversight and international cooperation.

The purpose of the future Framework Convention is set out in Article 1, namely 'to ensure that activities within the lifecycle of artificial intelligence systems are fully consistent with human rights, democracy and the rule of law', thereby addressing the entirety of the system's lifecycle and emphasising the human-rights based approach to AI governance. Considering that the instrument may not directly confer rights upon individuals, the signatories will be obliged to introduce measures to give effect to the provisions set out in the Convention.[104]

The Framework Convention utilises the term 'artificial intelligence system' which is defined in Article 2 as 'a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that may influence physical or virtual environments' in recognition that different 'systems vary in their levels of autonomy and adaptiveness after deployment'. Applying to the whole lifecycle of an AI system, the Draft Framework Convention adopts a risk-based approach to the design, development, and use of AI systems[105] to protect human rights[106] and ensure the integrity of democratic processes and respect for rule of law[107].

General common principles that the parties shall implement in regard to AI systems include measures to respect human dignity and individual autonomy,[108] to ensure adequate transparency and oversight,[109] to ensure accountability and responsibility for human rights violations,[110] to respect equality and non-discrimination,[111] to ensure privacy and personal data protection,[112] to preserve health [and probably the environment],[113] to promote reliability and trust,[114] and foster safe innovation[115]. In addition, the Convention sets out a requirement for remedies and procedural safeguards.[116]

In comparison to the January 2023 revised zero Draft, the Draft of December 2023 sets out a risk and impact management framework that requires the

99 Ibid, para 14.1.

100 Ibid, para 14.2.

101 Council of Europe, CAI, 'Revised Zero Draft [Convention] on Artificial Intelligence, Human Rights, Democracy and the Rule of Law' (n 29).

102 Council of Europe, CAI, 'Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law' (18 December 2023) CAI(2023)28 <https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043>.

103 David Leslie et al, 'Artificial Intelligence, Human Rights, Democracy and the Rule of Law, A Primer' (2021), 29 <https://rm.coe.int/primer-en-new-cover-pages-coe-english-compressed-2754-7186-0228-v-1/1680a2fd4a>.

104 Cf art 1(2) Draft Framework Convention.

105 Art 16 Draft Framework Convention, which sets out a risk and impact management framework.

106 Art 4 Draft Framework Convention.

107 Art 5 Draft Framework Convention.

108 Art 6 Draft Framework Convention.

109 Art 7 Draft Framework Convention.

110 Art 8 Draft Framework Convention.

111 Art 9 Draft Framework Convention.

112 Art 10 Draft Framework Convention.

113 Art 11 Draft Framework Convention.

114 Art 12 Draft Framework Convention.

115 Art 13 Draft Framework Convention.

116 Arts 14 and 15 Draft Framework Convention.

parties to take measures regarding the identification, assessment, prevention and mitigation of risks including inter alia a documentation of the risk and impact management process.[117] This also includes a mechanism for a moratorium or ban of certain AI uses.[118] The Draft Convention also emphasises the need for international co-operation to realise the purpose of the Convention.[119]

IV. Examples for Approaches on National Level

1. The US Presidential Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Starting in 2016 with the release of a first National AI R&D Strategic Plan,[120] the United States developed activities in the field of AI very early on.[121] Despite a strong bipartisan interest in AI regulation[122], disagreement remains on how to regulate AI at federal level.

An important milestone was achieved in 2022, with the release of the 'Blueprint for an AI Bill of Rights'[123] by the US White House Office of Science and Technology Policy which is accompanied by several related agency actions.[124] The non-regulatory and non-binding Blueprint lays out five principles that should guide the design, use and deployment of AI, namely: (1) safe and effective systems, (2) algorithmic discrimination protections, (3) data privacy, (4) notice and explanation, and (5) human alternatives, consideration and fallback.

Furthermore, the push for national AI standards through the executive branch has intensified in October 2023 with another federal government effort to guide responsible AI development, namely, the US 'Presidential Executive Order (E.O. 14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'[125]. The purpose of the Executive Order is to set out policy goals to support the development of responsible AI technologies by mitigating the substantial risks of AI to harness 'AI for good' and realise 'its myriad benefits'.

Content-wise, the Executive Order builds on the Blueprint for an AI Bill of Rights and the 'National Institute of Standards and Technology's AI Risk Management Framework'[126]. The Executive Order identifies eight overarching guiding principles and priorities: (1) ensuring safety and security,[127] (2) promoting innovation and competition,[128] (3) worker support,[129] (4) advancing equity and civil rights,[130] (5) consumer protection,[131] (6) privacy protection,[132] (7) advancing Federal Government use of AI,[133] and (8) strengthening international leadership[134]. Addressees of the Executive Order are executive departments and agencies which shall adhere to these principles.[135]

Similar to the OECD Recommendation and the Council of Europe Draft Framework Convention, the E.O. provides a definition of the term 'artificial intel-

117 Art 16 Draft Framework Convention.

118 Ibid.

119 Art 25 Draft Framework Convention.

120 Executive Office of the President of the United States, 'The National Artificial Intelligence Research and Development Strategic Plan' (October 2016) <https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf>.

121 For an overview of the US approach, see Huw Roberts et al, 'Achieving a 'Good AI Society': Comparing the Aims and Progress of the EU and the US' (2021) 27 Sci Eng Ethics Art no 68, p 11.

122 Eg SAFE Innovation S.4488 – Global Catastrophic Risk Management Act of 2022; Office of US Senator Ron Wyden, Algorithmic Accountability Act of 2022 (2022) 117th Congress 2D Session (rejected) <https://www.congress.gov/bill/117th-congress/senate-bill/3572>; S. Rept 117-254; Artificial Intelligence Research, Innovation, and Accountability Act of 2023.

123 The White House, 'Blueprint for an AI Bill of Rights' (October 2022) <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>.

124 For related actions see the White House, 'Fact Sheet: Biden-Harris Administration Announces Key Actions to Advance Tech Accountability and Protect the Rights of the American Public' (4 October 2022) <https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/>.

125 Executive Order (E.O.) 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence of 30 October 2023 <https://rb.gy/quetej>.

126 National Institute of Standards and Technology, 'Artificial Intelligence Risk Management Framework (NIST AI 100-1)' (January 2023) <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

127 S 4 E.O. 14110.

128 S 5 E.O. 14110.

129 S 6 E.O. 14110.

130 S 7 E.O. 14110.

131 S 8 E.O. 14110.

132 S 9 E.O. 14110.

133 S 10 E.O. 14110.

134 S 11 E.O. 14110.

135 Ibid.

ligence' and 'AI system', while also providing an extensive list of further definitions including definitions for 'AI model', 'dual-use foundation model', 'generative AI' or 'synthetic content'.[136] According to Section 3(b) the term 'artificial intelligence' has the meaning set forth in 15 U.S.C. 9401(3), 'a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action'.

The Executive Order directs more than 50 federal entities to implement the principles and priorities in more than 100 specific actions including the development of further guidelines, regulation and best practices. It must be noted that the policy and principles of the Executive Order have a market-driven focus and aim at protecting the interests of Americans; accordingly, the Executive Order states with regard to the interests of individuals that privacy and civil liberties of citizens must be protected.[137]

2. The UK White Paper Proposing a Pro-Innovation Approach to Regulation

In contrast to a cross-cutting AI-specific regulation, the UK follows a context-specific, sectoral approach to regulating AI which is addressed as 'a pro-innovation approach to AI regulation'. A keen interest in appropriate AI governance[138] is expressed in the National AI Strategy with the aim to establish the 'most trusted and pro-innovation system of AI governance in the world'.[139]

A 2018 House of Lords' report delineates the UK's approach to regulation, namely, that 'blanket AI-specific regulation, at this stage, would be inappropriate... [and] that existing sector-specific regulators are best placed to consider the impact on their sector of any subsequent regulation which may be needed'.[140] This decentralised, sector-lead approach to regulation was confirmed by the Government in June 2018[141] and reaffirmed in 2023 in the White Paper 'A pro-innovation Approach to AI Regulation'[142].

The AI White Paper suggests an agile principles-based framework for regulators to interpret and apply to AI within their remits, which is built around four key elements: (1) 'defining AI based on its unique characteristics to support regulator coordination',[143] (2) 'adopting the aforementioned context-specific approach',[144] (3) 'providing a set of cross-sectoral principles to guide regulator responses to AI risks and opportunities',[145] and (4) 'delivering new central functions to support regulators to deliver the AI regulatory framework, maximising the benefits of an iterative approach and ensuring that the framework is coherent'[146].

Arguing that there is no uniform definition of AI and that a rigid definition runs the risk of being outdated quickly, the White Paper identifies 'adaptivity' and 'autonomy' as the main two characteristics of AI, which generate the need for a regulatory response.[147] Thus instead of a general definition, an agile approach is used that refers to the functional capabilities to address the challenges created by the two characteristics.

Notably, the White Paper seeks to address the regulation of the use of AI and not the underlying technology as such. A set of five values-focused cross-sectoral principles is established that should be implemented by regulators when governing AI within their merits.[148] These principles build on and reflect the OECD values-based AI principles and consist of: safety, security and robustness; appropriate trans-

136 S 3 E.O. 14110.

137 See s 2(e) and (f) E.O. 14110.

138 Huw Roberts et al 'Artificial intelligence regulation in the United Kingdom: a path to good governance and global leadership?' (2023) 12(2) Internet Policy Review <https://policyreview.info/pdf/policyreview-2023-2-1709.pdf>.

139 UK Government, 'National AI Strategy' (updated 18 December 2023) <https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version>.

140 House of Lords Artificial Intelligence Select Committee, 'AI in the UK: Ready, willing and able" (2018) HL Paper 100, para 386 <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>.

141 UK Parliament, 'Government response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able" (2018) CM 9645, paras 103 et seq.

142 UK Government, Secretary of State for Science, Innovation and Technology, 'A Pro-Innovation Approach to AI Regulation' (Command Paper No 815, 2023).

143 Ibid, s 3.2.1.

144 Ibid, s 3.2.2.

145 Ibid, s 3.2.3.

146 Ibid, s 3.2.4.

147 Ibid, s 3.2.1.

148 See also Roberts et al (n 138), 8.

parency and explainability; fairness; accountability and governance; and contestability and redress.[149]

It is envisaged that legal responsibility for compliance with the principles is allocated to the actors in the AI life-cycle that are best placed to identify, assess and mitigate AI risks effectively.[150] In view of the emergence of foundation models, the White Paper stresses the commitment for an adaptable, proportionate approach that considers the issues raised by these powerful models without detailing a concrete approach.

The White Paper opened up for consultation and in February 2024, the UK Department of Science, Innovation and Technology, published a Government response to this consultation.[151] The Response reaffirms the context-based approach which avoids blanket rules applying to AI technology. A roadmap enlists the next steps to be taken in 2024 including the support of international collaboration on AI governance.[152]

V. The European Union's AI Act

In April 2018, the European Commission's Communication 'Artificial Intelligence for Europe' set out a European initiative on AI, which inter alia should aim to ensure an appropriate ethical and legal framework based on the Union's values and respecting fundamental rights.[153] In December 2018, the European Commission presented its Coordinated Plan on Artificial Intelligence[154] which foresaw inter alia the development of AI ethics guidelines[155] and the monitoring of existing[156] and emerging[157] legal frameworks in view of their application in the context of AI. In the following, Ethics Guidelines for Trustworthy AI were published in April 2019 which included a 'Framework for Trustworthy AI'.[158]

Following a European Commission White Paper on AI[159] in February 2020, the European Commission presented its AI package which included a Communication on fostering a European approach to AI[160], a review of the Coordinated Plan on Artificial Intelligence[161] and a Proposal for a 'Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts'[162] in April 2021.

The Proposal for an AI Act was complemented by a number of legislative proposals in the area of product safety as well as a Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)[163] and a Proposal for a Directive on liability for defective products[164] of 28 September 2022. The latter seeks to ensure that when AI systems are defective and cause physical harm, property damage or data loss to seek

149 UK Government, Secretary of State for Science, Innovation and Technology, 'A Pro-Innovation Approach to AI Regulation' (n 142), s 3.2.3.

150 Ibid, s 3.3.2.

151 UK Government, 'A Pro-Innovation Approach to AI Regulation, Government Response to Consultation' (February 2024) <https://assets.publishing.service.gov.uk/media/65c1e399c43191000d1a45f4/a-pro-innovation-approach-to-ai-regulation-amended-governement-response-web-ready.pdf>.

152 Ibid, para 93.

153 European Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe', COM(2018) 237 final, 4.

154 European Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Coordinated Plan on Artificial Intelligence' COM(2018) 795 final, revised in 2021.

155 This has been transposed by the High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for trustworthy AI' (2019) <https://rb.gy/vfm303>.

156 Such as the EU safety framework including the Product Liability Directive and the Machinery Directive.

157 Such as the GDPR and Proposals under the Digital Single Market Strategy.

158 High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for trustworthy AI' (n 155), 6 et seq. For a summary of the Framework for Trustworthy AI, see Bart van der Sloot, *Regulating the Synthetic Society* (Hart Publishing 2024), 144 et seq.

159 European Commission, 'White Paper on Artificial Intelligence – A European approach to excellence and trust' COM(2020) 65 final.

160 European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Fostering a European approach to Artificial Intelligence' COM(2021) 205 final.

161 European Commission, 'Annexes to the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Fostering a European approach to Artificial Intelligence' COM(2021) 205 final

162 Commission, 'Proposal for a Regulation laying down harmonized rules on artificial intelligence' (n 12).

163 European Commission, 'Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)' COM(2022) 496 final.

164 European Commission, 'Proposal for a Directive of the European Parliament and of the Council on liability for defective products' COM(2022) 495 final.

compensation from the AI system provider or from any manufacturer that integrates an AI system into another product.

In the context of the ordinary legislative procedure, numerous amendments have been proposed to the Proposal for an AI Act by the Council and European Parliament alike.[165] Following a fifth trilogue meeting in December 2023, the interinstitutional negotiations between the co-legislators closed and the Belgian presidency of the Council of EU Ministers presented on 24 January 2024 the final version of the text at a technical meeting.[166]

As a Regulation, the AI Act will have binding force in all EU Member States, rendering the Act a forerunner in AI regulation as hard law. Besides the market harmonisation aspect, the Act seeks to promote a human centric and trustworthy AI, while ensuring a high level of protection of health, safety and fundamental rights.[167] The AI Act[168] introduces rights and obligations for providers, deployers, importers and distributors, product manufacturers, authorised representatives and affected persons located in the EU.[169] This scope of application is based on an extensive understanding of the market principle which may lead to an externalisation of norms and standards when Article 2(1) lit. c) links the application of the AI Act to the output of an AI system used in the EU.[170]

The negotiations included a number of substantial changes to the originally proposed definition of an AI system, which had been widely criticised[171] and subsequently underwent closer alignment with the definition of the OECD. According to Article 3(1) an 'AI system' is now defined as 'a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.

As regards the potentially wide scope of this definition, Recital 6 clarifies that the definition is not intended to cover simpler traditional software systems or programming approaches. Furthermore, the AI Act also provides a definition for general purpose AI (GPAI) model and GPAI systems, whereby GPAI model means 'an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications'.[172]

The AI Act follows a risk-based approach[173] and distinguishes between prohibited AI practices,[174] high-risk AI systems,[175] and AI systems other than high-risk AI systems[176]. While non-prohibited AI systems have to respect certain transparency obliga-

165 For an analysis of the original Commission Proposal see Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act – Analysing the Good, the Bad and the unclear Elements of the Proposed Approach' (2021) 22 CRi 97. For a summary of the AI Act following Council amendments see Lilian Edwards, 'The EU AI Act: A Summary of its Significance and Scope' (Ada Lovelace Institute, 2022) <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf>, and Emre Kazim et al, 'Proposed EU AI Act – Presidency Compromise Text: Select Overview and Comment on the Changes to the Proposed Regulation' (2023) 3 AI and Ethics 381.

166 Council of the European Union, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Analysis of the final compromise text with a view to agreement' (26 January 2024), Interinstitutional file 2021/0106(COD) 5662/24 <https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf>. For a legal analysis of the final draft, see David Bomhard and Jonas Siglmüller, 'AI Act – das Trilogergebnis' (2024) RDi 45. See also for a detailed overview of the outcome Marinos Kalpakos, 'Defining the Future: The AI Act's Potential in equitably Safeguarding Fundamental Rights and Promoting AI Innovation', UFITA 2023 (1), forthcoming.

167 Art 1(1) AI Act (compromise agreement of 26 January 2024).

168 All references made to the AI Act – unless otherwise stated – refer to the compromise agreement of 26 January 2024.

169 Art 2(1) AI Act.

170 Cf Bomhard and Siglmüller (n 166), 46.

171 Hannah Ruschemeier, 'AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal' (2023) 23 ERA Forum, 361–376 <https://doi.org/10.1007/s12027-022-00725-6>; David Bomhard and Marieke Merkle, 'Europäische KI-Verordnung, Der aktuelle Kommissionsentwurf und praktische Auswirkungen' (2021), 276, 277; Hans Steege, 'Definition von Künstlicher Intelligenz in Art 3 Nr. 1 KI-VO-E – Ein Meilenstein auf dem Weg zu einem harmonisierten Rechtsrahmen?' (2022) MMR 926, 927; Andreas Ebert and Indra Spiecker gen. Döhmann, 'Der Kommissionsentwurf für eine KI-Verordnung der EU – Die EU als Trendsetter weltweiter Regulierung' (2021) NVwZ 1188, 1189; Gerald Spindler, 'Der Vorschlag der EU-Kommission für eine Verordnung zur Regulierung der Künstlichen Intelligenz (KI-VO-E)' (2021) CR 361, 363.

172 Art 2(44b) AI Act; this does not cover AI models that are used before release on the market for research, development and prototyping activities. For the definition of GPAI system see Art 2(44e) AI Act.

173 For the pitfalls of this approach see van der Sloot (n 158), 153.

174 Art 5 AI Act.

175 Arts 6 et seq AI Act.

176 Art 69 AI Act.

tions[177], non-high-risk face no further obligations but are encouraged to draw up of codes of conduct concerning the voluntary application of specific requirements. Requirements for high-risk AI systems relate to risk management system,[178] data and data governance,[179] technical documentation,[180] record-keeping,[181] transparency and information sharing,[182] human oversight,[183]as well as accuracy, robustness and cybersecurity.[184] The obligations for providers and deployers of high-risk systems draw on the model for creating safe and secure products under the New Legislative Framework (NLF)[185]. Accordingly, there is strong reliance on private self-regulation including approved conformity assessments with codes of practice and 'alternative adequate means'.[186]

Following the launch of OpenAI's GPT-4, amendments had been introduced to address also GPAI models.[187] The compromise agreement now includes horizontal obligations for all GPAI models which include inter alia technical documentation obligations. Additional requirements such as risk assessment and mitigation measures exist for models with so called 'systemic risks' which derive from their advanced capabilities.

For GPAI models a new governance structure is introduced with an AI Office, whereas to other AI systems a market surveillance system applies at national level. Furthermore, a European AI Board (EUAIB) will be established to advise and assist the Commission and Member States in the consistent and effective application of the AI Act.[188]

The AI Act also contains measures in support of innovation, which include inter alia the establishment of regulation sandboxes at national level.[189]

VI. A Joint Consideration of Policies and Regulatory Initiatives: Common Developments and Overall Progress

This overview of emerging instruments addressing AI shows that regulation of AI is evolving quickly, and that the speed of regulatory and policy responses accelerates with increasing capabilities of AI technologies.

The main challenge for regulation is posed by the fact that AI is not a product that once put on the market does not further evolve during its lifecycle. Moreover, machine learning technologies learn from data, which means that known and unknown challenges in how these systems act are introduced, posing new risk profiles.[190] Furthermore, while within other domains of engineering this may be the case, established technology readiness level assessment methods are not applicable to AI.[191] As a result of the particularities of the AI lifecycle, many regulatory instruments addressing AI refrain from a static definition of AI and AI system.

Further, attributing responsibility to certain actors involved in the different phases of the AI lifecycle is equally challenging, meaning that as a means of minimum consensus high-level principles are agreed at global level rather than attributing certain obliga-

177 Art 52 AI Act.

178 Art 9 AI Act.

179 Art 10 AI Act.

180 Art 11 AI Act.

181 Art 12 AI Act.

182 Art 13 AI Act.

183 Art 14 AI Act.

184 Art 15 AI Act.

185 The NLF seeks to ensure the safety of products entering and circulating in the internal market and consists of Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93 [2008] OJ L218/30, the Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC [2008] OJ L218/82, and Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011 [2019] OJ L169/1.

186 Cf in that regard recital 60t and art 52c(3) AI Act.

187 For the genesis of rules for GPAI models, see Bomhard and Siglmüller (n 166), 49.

188 Arts 57 et seq AI Act.

189 Arts 53 et seq AI Act. In that regard, it has already been argued that the provisions do not pay sufficient regard to SMEs and OSS providers, see Bomhard and Siglmüller (n 166), 54.

190 Alexander Lavin et al, 'Technology Readiness Levels for Machine Learning Systems' (2022) 13 Nature Communications Art no 6039, 1. See also Lilian Edwards, 'Regulating AI in Europe: Four Problems and Four Solutions' (Ada Lovelace Institute, 2022), 6 <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf>.

191 Cf Lavin et al (ibid).

tions to specific actors. Thus, the latter would go too far in view of intergovernmental policy statements. In this respect, the Council of Europe Draft Framework Convention remains very general requiring parties to 'take measures necessary to ensure accountability, responsibility and legal liability [...] resulting from the application of' AI systems.[192]

What can be seen is that from a first focus on ethical AI initiatives with the recognition of 'aspirational principles'[193] of fairness, accuracy, transparency that address the underlying technology, recent approaches also seek to ensure and promote innovation and investment in innovation. Obviously, there is a thin line between regulation and over-regulation that might hinder innovation and investment and result in lagging behind technologically.

Developing the principles-based approach enshrined in the OECD Recommendation and replicated by the G7 and G20 further, the UNESCO instrument evolved to also include governance aims regarding the usage of the output in terms of inter alia gender equality and within the employment context. Where certain usage is however limited or banned, an ongoing challenge is the ability to enforce such bans globally.[194]

Further progress in AI policies and regulation was certainly triggered by the new challenges posed by GPAI systems. In fact, the AI Safety Summit and the resulting Bletchley Declaration as well as the US Presidential Executive Order are a direct response to the increased capabilities of GPAI. Notably, a compromise on the EU AI Act was on the brink at the last minute with disagreement on how GPAI should be regulated. In the incoming EU AI Act, similar to AI systems in general, GPAI models also face a graduated approach depending on whether they are subject to 'systemic risks' stemming from their capabilities mainly in computing power.

With the EU AI Act, there will not only be a binding instrument in place, that addresses GPAI and AI systems horizontally, but also an instrument that applies to providers and deployers from outside the EU whose outputs are accessible to or used in the EU and thereby has extraterritorial reach. Legislation in the field of digital economy naturally influences global markets when drafted by an important market for data-driven businesses.

This externalisation of EU law has been referred to as the 'Brussels Effect'[195] where in particular the data protection regime under the GDPR has proven a strong example of said effect. Whether the AI Act will shape international standards in the same way as the GDPR or whether stronger players in the AI market, notably the US with its economic weight in tech leadership, will set standards, remains to be seen.[196] So far, the EU is taking the lead in regulation, while US efforts are modest and remain rather fragmented. In consideration of sectoral approaches to regulation, the impact of EU regulation with its risk-based approach and focus on mandatory self-regulation for many applications is also going to vary. Much of the success of the AI Act and any kind of AI regulation will depend on its implementation, the ability to balance potentially contravening interests, and ultimately, in view of the pace of technological progress, on the ability to adapt. This will necessitate a continuous and critical observation and assessment of regulatory approaches and their application, which is the reason for creating a new journal and which will be the core mission of AIRe.

---

192 Art 14 Draft Framework Convention.

193 Karen Yeung, 'Introductory Note to Recommendation of the Council on Artificial Intelligence (OECD)' (2020) 59 International Legal Materials 27.

194 For instance, China, having endorsed the UNESCO Recommendation, still utilises social scoring which is a use of AI that is so intrusive that it even affects human dignity. See also Rotenberg (n 72), 619.

195 Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (OUP 2019), 142.

196 For a de jure Brussels Effect of parts of the AI Act see Charlotte Siegmann and Markus Anderljung, 'The Brussels Effect and Artificial Intelligence, How EU Regulation Will Impact the Global AI Market' (Centre for the Governance of AI, 2022) <https://arxiv.org/ftp/arxiv/papers/2208/2208.12645.pdf>; see also Jakob Mökander et al, 'The US Algorithmic Accountability of 2022 vs. the EU Artificial Intelligence Act: What Can they Learn from Each Other?' (2022) 32 Minds and Machines 751, 755.

International: Annex

AI Regulation and Governance on a Global Scale: Overview of Scope, Definitions and Key Elements

The table below serves as annex to the report 'AI Regulation and Governance on a Global Scale'. It outlines in a summarised form and as synoptical comparison current international, regional and examples for national regulatory initiatives aimed at regulation and governance of AI systems. The table was prepared by *Dr Sandra Schmitz-Berndt*, Research Associate in the project LAIWYERS, funded by the Institute of Advanced Studies (IAS) of the University of Luxembourg (for correspondence: <Sandra .schmitz@uni.lu>).

	OECD	**UNESCO**	**CoE**	**US**	**UK**	**EU**
Year	2019, updated 2023	2021	Forthcoming (planned 2024)	2023	2023	2024
Instrument	Recommendation of the Council on Artificial Intelligence	Recommendation on the Ethics of Artificial Intelligence	Draft for a Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law	Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence	White Paper 'A pro-innovation approach to AI regulation'	Regulation laying down harmonised rules on AI (AI Act)
Binding/ Non-Binding	Non-binding	Non-binding	Non-binding per se / Binding for Signatories	Binding	Non-binding	Binding
Objectives of Instrument	Foster innovation and trust in AI by promoting the responsible stewardship of trustworthy AI	Make AI systems work for the good of humanity, individuals, societies and the environment and ecosystems, and to prevent harm. It also aims at stimulating the peaceful use of AI systems	Ensure that activities within the lifecycle of AI systems are fully consistent with human rights, democracy and the rule of law	Ensure the safe, secure and trustworthy development and use of AI	Pro-innovation approach to AI regulation	Improve the functioning of the single market, promote the uptake of human-centric and trustworthy AI, while ensuring a high level of protection of health, safety and fundamental rights, and to support innovation

	OECD	UNESCO	CoE	US	UK	EU
Scope of application	Adherents to the Recommendation and all AI actors	Member States and all AI actors	Activities by public or private entities undertaken within the AI lifecycle that have the potential to interfere with human rights, democracy and the rule of law	Executive departments and agencies undertaking the actions set forth in the order	Mainly regulators, indirectly industry and individual businesses developing and using AI	Providers, deployers, importers and distributors of AI systems and authorised representatives of providers not established in the EU; affected persons located in the EU
AI Definition	An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment	No fixed definition. AI systems are considered as systems which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control	An AI system is a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that may influence physical or virtual environments. Different artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment	The term AI system means any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI. Further definitions for 'AI model', 'dual-use foundation model', 'generative AI' or 'synthetic content'	No fixed definition. Agile approach to encompass products and services that are 'adaptable' and 'autonomous'	An AI system is a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments
General Approach	Value-based principles: (1) Principles for responsible stewardship of trustworthy AI, (2) Recommendations to policy-makers relating to national policies and international cooperation	Value-based principles: (1) Complementary and interrelated values and principles, (2) Identification of areas of policy actions, including ethical impact assessment, and international cooperation	Value-based principles: (1) General obligations, (2) Principles related to activities within the lifecycle of AI systems, (3) Remedies, (4) Obligations regarding the assessment and mitigation of risks (5) Guidance to ensure oversight and international cooperation	Principles-based: Eight guiding principles and priorities	Principles-based framework building on the OECD principles	Product safety and risk-based approach with four levels of risk: Unacceptable risk, high risk, limited risk, minimal or no risk and separate category of general purpose AI