AI governance ethics : artificial intelligence with shared values and rules

Global 22

AI Governance Ethics

Artificial Intelligence with Shared Values and Rules

Christoph Stückelberger, Maria Merchán Rocamora, Divya Singh, Pavan Duggal (Eds.)

Globethics

AI Governance Ethics

Artificial Intelligence

with Shared Values and Rules

AI Governance Ethics

Artificial Intelligence
with Shared Values and Rules

Christoph Stückelberger, María Merchán Rocamora,
Divya Singh, Pavan Duggal (Editors)

Globethics Global Series No. 22

Globethics Global Series

Christoph Stückelberger, Founder and President of Globethics and Professor of Ethics (emeritus University of Basel/ Switzerland, Visiting Professor in Enugu/Nigeria, Beijing/China, Moscow/Russia, Leeds/UK).
Fadi Daou, Executive Director of Globethics and Professor of Religious Studies

Globethics Global Series 22
Christoph Stückelberger, María Merchán Rocamora, Divya Singh, Pavan Duggal (Eds.), AI Governance Ethics. Artificial Intelligence with Shared Values and Rules
Geneva: Globethics, 2024
DOI: 10.58863/20.500.12424/4318987
ISBN 978-2-88931-609-0 (online version)
ISBN 978-2-88931-610-6 (print version)

Managing Editor: Dr Ignace Haaz
Assistant Editor: Jakob W. Bühlmann
Cover design: Michael Cagnoni

Chemin du Pavillon 2
CH 1218 Le Grand-Saconnex, Geneva
Switzerland
Website: *https://globethics.net/publications*
Email: *publications@globethics.net*

All web links in this text have been verified as of October 2024.
The electronic version of this book can be downloaded for free from the Globethics website: https://globethics.net

TABLE OF CONTENTS

Part III

Public Sector AI Guidelines, Standards, Regulations and Ethics: Global, Continental, National

Part IV

Private Sector AI Guidelines, Standards Regulations and Ethics: Global

0

INTRODUCTION

Christoph Stückelberger, María Merchán Rocamora,
Divya Singh, Pavan Duggal (Editors)

Is there a relationship between Artificial Intelligence (AI) and ethics, and if so, what is it, where does it play out, and how do we ensure its application? How does the governance of AI relate to ethics? Can we expect the governing frameworks to keep abreast in this fast-paced environment, where the *in-house* capabilities of regulatory bodies appear to be so far behind those of a cluster of industry leaders? Ever since the creation of the International Telecommunications Union in 1865, technological developments have pushed national boundaries by reaching a universality in adoption that contrasts with typical regulatory models and the variety of approaches by which their use is framed.

01 Who governs whom with which values?

When tackling the edition of a volume on *global governance of AI*, we must contend with the many open questions about *whose values* (and moreover, *whose interests*) are at stake. On AI Ethics, a convergence of

values and criteria for ethical AI can be observed in the academic, public, private, business, and civil society sectors.[1]

In a moment characterised by ever greater velocity and an evident urgency, regulators try to catch up through executive orders that set concrete timeframes (such as the 2023 White House's Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[2]), or speed up legislative processes: The EU AI Act[3] was approved within two years, whereas six years went by between the proposal and enforcement of the General Data Protection Regulation (GDPR)[4].

Currently we are seeing an emerging pattern that appears to coalesce from multiple data points as if resulting from an unsupervised machine-learning algorithm. The internal protocols of companies continue to evoke principles of non-discrimination, explainability[5], data protection and accountability, with a recurrent emphasis on "human-centered AI

[1] See e.g. Kristian McCann, *Top 10 Ethical AI Considerations,* Aio magazine, 26 June 2024. https://aimagazine.com/articles/top-10-ethical-considerations%0A

[2] See in this book below, chapter 21.

[3] See in this book below, chapter 18.

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation): https://eur-lex.europa.eu/eli/reg/2016/679/oj

[5] The concept of "explainability", very commonly referred to when it comes to discussing AI regulation and safeguards, alludes to the ability to understand and interpret how an AI system makes its decisions. For a model to be *explainable*, its decision-making process -especially those based on machine learning- should ideally be describable and interpretable even for non-experts.

development".[6] The widely adopted UNESCO Recommendation on AI ethics[7] reflected the same language at its adoption in 2021 by the 193 Member States of UNESCO's General Conference. In the AI EU Act, similar core sentiments and themes are repeated, grounded at least in part in treaty-based human rights commitments. However, across the board, real commitment (beyond the paper promises) and enforcement capacities remain a challenge for many of the regulators.

Regulators are increasingly seeking companies' feedback, in an attempt to include state-of-the-art knowledge into their considerations and under the recurrent threat of "stifling innovation". Even the European Union, often recognised by the public as less lenient towards corporate interest, after taking the legislative lead, launched a consultation with Google experts in September 2024. The topic under discussion was a reflection on how to protect one of the core features of the EU's internal market namely, antitrust[8]. Leading companies have developed detailed self-assessments, making pledges to self-regulate responsibly. This might mean that they are "*essentially writing the exam by which they are evaluated*", to quote Brandie Nonnecke (Founding Director of the CITRIS Policy Lab at UC Berkeley)[9], in a context where the *evaluators* might lack

---

[6] See in this book below, Part IV, chapters 23-27.

[7] UNESCO's Recommendation on the Ethics of Artificial Intelligence, adopted on November 25, 2021: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence. On Unesco, see also below, chapters 15 and 16.

[8] "EU regulators to seek feedback on Google's compliance proposals to avert charges", Reuters: https://www.reuters.com/technology/eu-regulators-seek-feedback-googles-compliance-proposals-avert-charges-2024-09-04/". (Accessed 30 Sept. 2024).

[9] Melissa Heikkilä, "AI Companies Promised the White House to Self-Regulate One Year Ago. What's Changed?", MIT Technology Review, July 22, 2024: https://www.technologyreview.com/2024/07/22/1095193/ai-companies-promised-the-white-house-to-self-regulate-one-year-ago-whats-changed/

the understanding of the very parameters being laid before their eyes. Yet, in absence of meaningful engagement from the public sector and the general public we end up facing the old paradox "*Quis custodiet ipsos custodes?*[10]". Only this time, the effective *watching* is even more difficult to be exercised as the prerogative of any one national government.

02 Globethics engagement for AI ethics

Globethics' Policy Report "Inclusive AI for a Better Future[11]*"* noted the need for a *multistakeholder* Geneva Compact for inclusive AI development and use, as a result of a series of policy dialogues conducted during the Geneva Peace Week in 2023. Globethics states:

> Due to the complex nature of AI, the need for a multistakeholder Compact for ethical reframing of the technology and its usage is formulated. The Compact aims to present the ethical framework in the most inclusive and simple way, for global people's by-in and legitimacy. It should allow the integration of different views and expectations, especially for those working in silos. The Compact needs to be intergenerational, coping with Children's Rights too ... The global multistakeholder Compact on Ethics of AI is more seen as a dynamic document, offering the ethical framing but not definitively defining neither the questions nor the answers. ... Recognizing that AI is both part of the shared challenges and the common good of humanity, the aim is to cultivate a continuous and inclusive policy engagement that advances collective action and convergent understanding."[12]

---

[10] From Juvenal's Satire VI, typically translated as "Who watches the watchmen?"
[11] Globethics, *Inclusive AI for a better Future. Policy Dialogue Report*, Geneva: Globethics Publications, 2024. Free download: www.globethics.net/publications.
[12] Ibid, 24-25.

In connection to Globethics' publications on *Cyber Ethics 4.0*[13] and *Data Ethics*[14] (the realm and the fuel of generative Artificial Intelligence), the Policy Report signalled how academia, the private and the public sectors have identified what the issues *are* to a great degree of convergence. The remaining questions are *how* to address them, and for *how long* they would remain the main quandaries. There are also competing priorities, even within the same set of norms and regardless of their degree of enforceability. On 22 September 2024, the United Nations adopted the *Pact for the Future*[15] during the Summit of the Future in New York. The Pact includes the *Global Digital Compact*[16], 16 pages consolidated after two years of consultations. How effective and representative these consultations are will remain of the essence, in a constant balancing act between stakeholders.

For instance, the Global Digital Compact calls for an *open* digital future for all. Debates about open source AI - what it actually is, and whether it is a potential threat for national security versus the advantages of increasing explainability and accessibility of AI models - requires involving a degree of cross-cutting, multisectoral and ethical, values-driven expertise that is difficult, but essential, to crowd in. Workers whose shifts are time-managed by an automated system might opt for explainability and accountability, while some security experts caution about the

---

[13] Globethics, *Cyber Ethics 4.0- Serving Humanity with Values,* Geneva: Globethics.net, 2018, Global Series No. 17. Free download: https://globethics.net/publications/cyber-ethics-40-serving-humanity-values

[14] Globethics, *Data Ethics: Building Trust. How digital technologies can serve humanity.* Geneva: Globethics, 2023. Global Series No. 18. Free download: https://repository.globethics.net/handle/20.500.12424/4273108

[15] Accessible at: https://www.un.org/sites/un2.un.org/files/sotf-the-pact-for-the-future.pdf

[16] Accessible at: https://www.un.org/global-digital-compact/sites/default/files/2024-09/Global%20Digital%20Compact%20-%20English_0.pdf

possibility of anyone downloading potentially harmful programming. Not all developers are *ad idem*, and their views are often dependant on what and how much they stand to lose in terms of strategic advantage. The capacity to enforce whatever these decisions are is another key aspect to consider.

03 This book: Contributions from all continents

In this volume, we have aimed to present a cross-cutting, geographically representative collection of articles, ethical guidelines, protocols, national and regional regulations that can support the reader in the direction of responsible AI in its development, use, deployment, and regulation. While many of these documents reflect the most recent developments at the time of publication, the febrile activity setting the tone in matters of AI governance is likely to produce new outcomes as we read. Representatives of China and the United States met privately in Geneva in May 2024 to discuss AI risks. The meeting indicated that the need for international communication (if not always direct cooperation) remains especially when stakes are high.[17] The United States and the European Union have respectively issued *Executive Order* 14110 and the *EU AI Act*. China passed the *Interim Measures for the Management of Generative Artificial Intelligence Services*, which came into force in August 2023.[18] These promulgations are very likely to be followed by other instruments in short order. The African Union also came up with an *African Global Compact* in 2024 and published its *Continental Artificial Intelligence*

[17] Michael Martina and Trevor Hunnicutt. "U.S. and China meet in Geneva to discuss AI risks," Reuters, May 13, 2024, https://www.reuters.com/technology/us-china-meet-geneva-discuss-ai-risks-2024-05-13/

[18] https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-china. (Accessed 30 Sept. 2024).

Strategy in July of this year (2024).[19] These recent developments show that all continents and sectors are running to catch opportunities and mitigate risks of AI.

Algorithm bias, lack of accountability, transformation of economic sectors, limitations placed by the energy consumption of data centres and the stark reality that we still face a global digital divide make questions around "human-centric AI" and "human rights-based AI" undeniable topical. However, it will be crucial for the public sector and the general public to develop a sound understanding of what these technologies are, what they require, and what they *can* do for productive contributions to be made in terms of safeguards, remedies and responsible adoption. Recognising the tsunami of developments in the field, being swift in accepting transversal talent and new competencies will be crucial for regulatory bodies to render constructive outcomes. Within lies the challenge to create dynamic settings, more responsive to shifting instruments – a task that requires a degree of interpretation, bridging the divide between technical language and that of regulatory frameworks. Failure to do so will result in stale or outright counterproductive outcomes.

This volume opens by providing a series of global ethical, legal, technical and multidisciplinary perspectives on AI regulation. In this rapidly changing environment, we consider here whether our existing understanding and applicability of human rights, ethics and values continues to hold firm at its core. The discussion then progresses to selected country-specific approaches, and closes with a collection of the main normative texts ("hard" and "soft" norms, from the public and private sectors), offering a comprehensive overview of the current AI governance landscape and the ethical considerations that underpin it. We are in the midst of

[19] Accessible here: https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy

transformational public debates including technofeudalism[20] and misinformation warnings, coupled with hopeful stories about better diagnostics, enhanced learning opportunities assisted by AI language models, increased efficiencies, and safer infrastructures. In this context, it is an ethical calling to spread digital literacy and enhance the competencies and capabilities across all sectors of society regardless of geographic location. To seek knowledge and - whatever the word is worth - the *truth* as we engage with the new normal of technology and specifically AI, that it might be a shared basis to develop ethical values sustaining the nascent architecture of the global governance of AI. Here is our modest contribution to this end.

Christoph Stückelberger, María Merchán Rocamora,
Divya Singh, Pavan Duggal (Editors)
Geneva/ Cape Town / New Delhi, 1 October 2024

[20] The term, popularised by Yanis Varoufakis 2023 book *"Technofeudalism: What killed capitalism"*, describes the modern digital economy as a new feudal system in which the oligopoly of large technological companies ("Tech giants") would control resources as feudal lords used, with users of the technology becoming dependent vassals or serfs- producing data without receiving economic compensation for it-; the precarious situation of *gig economy* workers is also compared to the conditions of medieval serfs, with little ability to influence the power relations they become dependent on.

PART I

AI GOVERNANCE: GLOBAL PERSPECTIVES

1

ETHICS IN THE CO-DRIVER SEAT OR ON THE EMERGENCY BRAKE?

ETHICAL CONSIDERATIONS ON GOVERNANCE INSTRUMENTS

Christoph Stückelberger, Switzerland

What is the role of ethics[21] in regulating and governing artificial intelligence? Are ethicists on the emergency brake? Are they in the co-driver seat? Are they in the commenting back seat or in the developers' lab? What are the phases in technological development where ethics comes in? What are ethical considerations in the spectrum of AI regulations? Before we deal with these questions let us start with some terminological clarifications.

1.1 Terminology: AI – Governance – Ethics

Artificial Intelligence (AI) is defined in very different ways. At the core, it is intelligence provided by computer systems and algorithms which can complement or replace human intelligence by digesting a huge

[21] *Christoph Stückelberger* is Professor of global ethics, emeritus at Basel University/Switzerland, Founder and Honorary President of Globethics Foundation, visiting professor in Nigeria, China and others.

amount of data. Manifold definitions are used in this book and in current debates. Critical is to distinguish a) between AI and machine-learning; and b) between generative and non-generative AI.

Governance means systems to direct, lead and control activities of humans or machines in order to increase efficiency and reduce risks. Governance is the broad term for different systems of governing and steering a (technological, ethical, economic, social) process.

AI Regulation means in the narrow sense the development of public policies and laws by legislators, in the broader sense it includes also voluntary standards and other more or less binding standards. AI regulations are the main instruments of governance of AI in this article.

Morale or morality is the set of values present in individuals in our society based on tradition and convention.

Ethics is the critical reflection of values and virtues which should guide human decisions and behaviour. Ethics can confirm existing moral values or modify them in light of a different value system or new societal developments.

AI Governance Ethics therefore aims at developing value-based orientations and benchmarks to govern and steer AI development.

1.2 Phases and speed of technological innovation

New technologies have been key in human development since thousands of years. Sharp natural silex stones have been used by humans as earliest stone toolkits at least 2.6 million years ago in the Early Stone Age[22] and used to cut meat and later making fire. Now we deal with AI

[22] Smithonian National Museum of National History: What does it mean to be human? https://humanorigins.si.edu/evidence/behavior/stone-tools/early-stone-age-tools Accessed 16 Aug 2024.

and future quantum technologies. An immense development. There are some phases in technological innovation which remain constant.

The first is the inspiration based on an innovative idea or an observation or a catastrophic experience. This inspiration often comes from any individual or a research group or develops over a long time by praxis, experience and the development of brain and skills as the silex example in the Stone Age shows.

The second phase of a technological innovation is testing using, improving for use by individuals or masses.

The third phase comes normally almost immediately with the second phase because the larger use needs a clarification of ownership and permission to use this technology and under which conditions. Developers want to protect their intellectual property, producers want to secure their future profit, consumer organisations want to protect consumers, military want to take it out from market to have it as unique technology for their defence etc.

On the history of AI[23]: The first phase of AI started around 1950 with research of computer scientists, with a rapid expansion in the 1960-1980es. The second phase of large implementation followed 1980-87 with a kind of AI timeout for lack of funding and then with new AI agents since the 1990ies to 2010. The second part of the second phase can be seen since 2012 with the AGI, the Artificial General Intelligence until the recent developments of OpenAI, chat CPT since 2020. Billions of dollars have been invested and the international, especially bipolar race US-China in AI development exploded. The third phase of regulatory efforts[24] in the last years since 2017 (when Elon Musk called for AI

[23] See e.g. https://www.tableau.com/data-insights/ai/history#ai-birth. Accessed 16 Aug 2024.

[24] Wikipedia, Regulation of artificial intelligence, https://en.wikipedia.org/wiki/Regulation_of_artificial_intelligence. Accessed 16 Aug 2024.

regulations) and an outline of coming implementations[25] shows that AI regulation is now the hot topic around the world.

The *speed of AI development* is one of the biggest ethical challenges of this new technology. In the Stone Age two million years ago, the first humans needed estimated about 500'000 years to develop first working tools from silex stones. Now, Chat GPT after its launch in December 2022, after one year was already used by a good part of humanity around the globe!

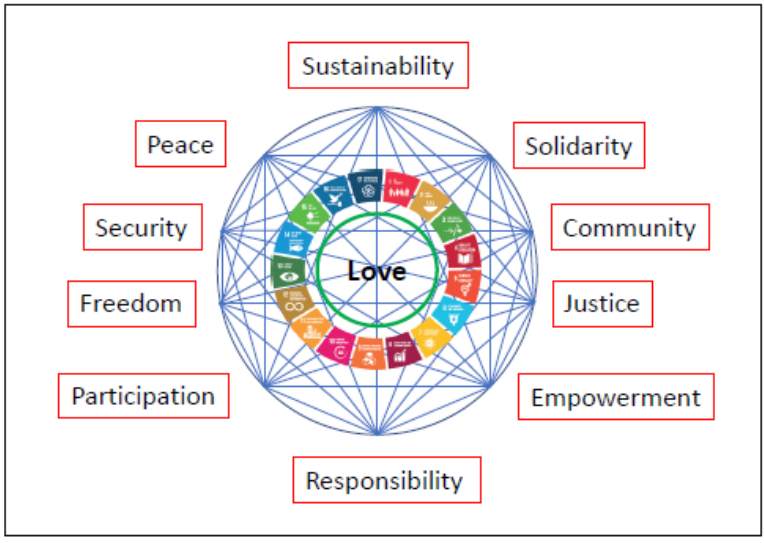
1.3 Values for ethical judgment of technological innovation

The anthropologically fundamental fact is, that humans are always able to do good and bad, life-supporting and life-destroying activities. That is the very essence of being human to have the freedom to decide for both whereas animals are – more or less- driven by their genes and instincts and therefore 'predetermined' in their actions. The freedom to decide leads to the question: which are the criteria to decide? This is the very beginning of ethics. I published extensively on values and ethical criteria; here I focus on a few short aspects.

Values are fundamental ethical benchmarks for decisions of individuals, groups, or institutions. Virtues are fundamental ethical benchmarks for individual character and behaviour. A list of ten values and virtues are mainly my value universe. Even though they are rooted in my Christian world view, I tested and developed them on all continents and can say, they are accepted throughout cultures and religions. The differences

[25] Nick Sherman, AI Regulations round the World, Mind foundry, Blog, 25 Jan 2024. https://www.mindfoundry.ai/blog/ai-regulations-around-the-world

come rather from the weight and priorities of specific values and virtues in specific contexts.[26]



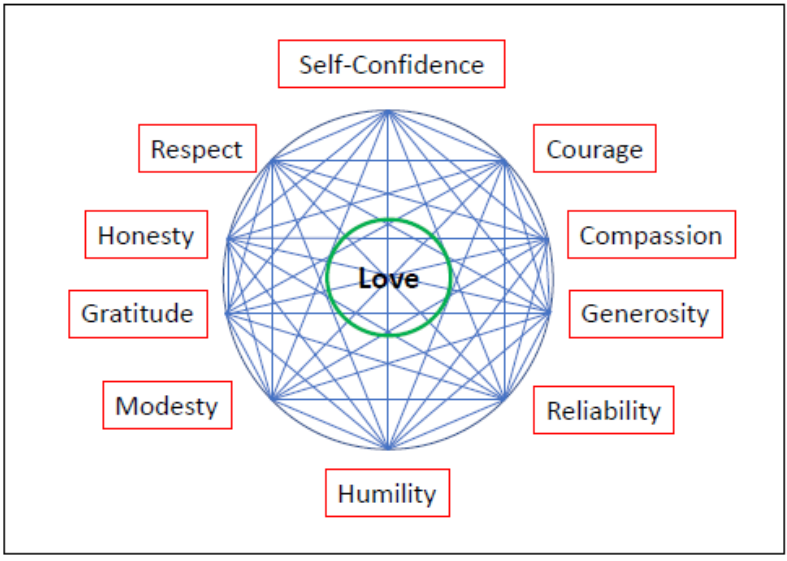
These ten values are fundamental for a life in dignity of individuals, of communities and for organisations on all levels. They also build the basis of the 17 UN Sustainable Development Goals ("SDGs"). They are in a circle related to each other. This shows the crucial need, not to isolate, but to connect and balance these values. This is the meaning of '*Globalance'*.

The same is the case for the following ten virtues. Love in the centre is then the holistic ability to act in a specific situation on the basis of a specific priority of a value or virtue for this specific situation while balancing and taking into account all other values or virtues.

---

[26] Graphs from Christoph Stückelberger, *Globalance towards a New World Order. Ethics Matters and Motivates. Handbook with 250 Graphs.* 2nd enlarged Edition, Globethics, Geneva, 2020, especially 237-272.

What do these values and virtues now mean for handling technologies in general, AI specifically and AI governance and regulations as the even more specific case of this article?



***Some selected examples of values-driven AI governance*:**

- *Freedom and Justice*: how can a free development and use of AI-tools be balanced with a fair, just access to its benefits?
- *Empowerment and Security*: How can users be empowered to use AI tools, while developers and regulators are meant to maintain high security standards jointly with them?
- *Peace and Responsibility*: How can AI tools be developed and implemented so that they serve peace and reduce the manyfold risks being abused for irresponsible, harmful activities e.g., in terror attacks or war?

1.4 The triangle of ethics, regulations and relations. Spectrum of governance instruments and their ethical relevance

To answer such difficult questions in relation to AI governance, we need to look at specific governance instruments. *The spectrum of AI governance instruments* is very large.

- It goes from very soft recommendations to very strong, legally binding laws, linked with hard sanctions in case of violation.
- The spectrum also includes levels of geographically binding character from local to national to continental and global.
- Governance instruments can be company specific, sector specific or cross-sectoral for any activity in the AI field.
- Governance instruments always have specific ethical, political, economic, and cultural contexts. What is understood as binding in one context, may be seen as wishful and a piece of paper. Even though, international law and conventions streamlined and unified common understanding of these instruments largely during the last hundred years.

Let us look at some regulatory instruments, their ethical importance, and weaknesses across both the public and private sectors:

- *Company internal rules:* They are the first step and needed as part of internal risk management and quality control within an organisation. An explicit value-orientation of a company/institution in its individual staff is part of the risk and reputation management! The strength of this instrument is the pro-active character of the company, which can be a Unique Selling Point and that it can be implemented in short time, the weakness of this instrument is that it may not correspond to sector standards or legislation.

- *Voluntary standards by industry sector or other sectors:* They are slower in development than internal rules, but faster than new legislations. The ethical strength is that they are based on intrinsic motivation of the sector, compared to compulsory legislation, the weakness is that they are voluntary and often relatively weak in content and enforcing instruments. The credibility depends very much on the sector and institutional settings. E.g., ISO standards are voluntary, but with a high credibility thanks to detailed monitoring mechanisms for each standard.
- *National binding legislations*: they are per definition legally binding within the particular country, enforceable by the judiciary and therefore ethically strong instruments. The credibility depends on strong enforcement mechanisms and a credible judiciary of the respective country. The ethical challenge is that it often takes a very long time to develop and implement new legislations (in Switzerland, a CO2-legislation was discussed over two decades). The ethical value also depends on the political system: an autocracy may decide a new AI regulation in a short time, but with a deficit of participation of key actors of the country in the development. In a democratic country, the process is slower, but the result normally better accepted and more sustainable in long term.
- *Continentally or regionally binding legislations*: it is ethically very similar to a national legislation. However, the ethical advantage is that the impact is higher as it is binding for many countries and allows better harmonised solutions, as e.g., the EU AI Act of 2024 shows. The price is that every member country e.g., such as EU members have to implement even though not every country may agree with the legislation.
- *Global conventions*: global, multilateral treaties e.g., from the UN and its sister organisations and agencies are ethically very positive as they

allow one standard or set of rules valid for all signatory states and therefore contributed to fair rules for everybody. The disadvantage is again that they often need much time to adopt and cannot be very strong as it needs consensus.

The broader governance instruments include:

- *Education to raise awareness*: Manyfold efforts to include AI education for children and students are booming. It is necessary to prevent from or at least risks of abuse from the side of developers, business, and users. As drivers of cars need a licence, a 'license' to use AI-tools may be the future, at the same time it is illusionary as far as AI is already everywhere in place. Ethics education needs to be part of AI education as benchmarks for what is good and what is harmful use of AI.
- *Databases on existing and planned regulations*: AI regulations are very fast developing, especially since the 2020s. Every month, new regulations are coming up. This book gives a glimpse and tries to offer some orientation is this jungle. However, databases on AI regulations can help to compare, adjust, and internationally coordinate AI regulations. Such publicly available databases are ethically very positive as contribution to participation and transparency.
- *Judiciary incl. courts* for implementing regulatory frames and sanctioning violators. Ethics without law is lame as we expressed above. The ethical quality of regulations depends on the strength of the enforcement regulations and the socio-political environment. A good law in a country where the judiciary is distorted by corruption, the good law is useless. The ethical strategy should then be to clean the courts from corruption before or while approving a new wave of regulations.

- *Media from mainstream to social media* are one of the priority concerns of AI misuse by 'fake news' (also referred to as mis- and disinformation) especially in social media and political (election) campaigns. Editorial standards, policies and media regulations are an important part of AI governance nowadays.
- *Investments* in AI and related companies are still the main driver in AI development, combined with geopolitical competition, especially US-China. The ethics of investors and ethical dialogue with investors and their view on ethical governance is crucial and one of the most powerful AI governance instruments.
- *Values-setting institutions* such as academic and religious institutions need to (continue to) develop values-based guidelines for the further development and use of AI. They are also critical sectors for education and training and serving as qualified dialogue partners. Globethics foundation is offering its expertise and multicultural academic network.
- *Databases on technical standards*: for the IT developers, programmers, industry, academics and regulators, the access to technical information on algorithms is needed in order to validate the chances and risks of such algorithms. The challenge is that most of this information is seen by the companies as their business secret and advantage in the market. The balance between the protection of intellectual property rights and public interest for transparency is a key ethical challenge and need.
- *Databases on risks analyses:* A new AI Risk Repository of MIT Future Tech in Boston/USA offers a living, searchable database of over 700 risks based on a taxonomy.[27] Every risk can be seen as an ethical

[27] What are the risks from Artificial Intelligence? A comprehensive living database of over 700 AI risks, https://airisk.mit.edu/ Accessed 16 Aug 2024.

challenge and chance. Therefore, this new database is also an impressive, helpful tool for ethical recommendations on AI governance and regulations.

The following **table A**[28] shows the causal taxonomy of AI risks. With my ethics eyes I can call it Causal taxonomy of AI Ethical challenges: risks by entity (humans, AI or others), risks by intention (intentional, unintentional, others) and as an important category risks in timing (pre-deployment, post-deployment, other).

Table A. Causal Taxonomy of AI Risks

Category	Level	Description
Entity	Human	The risk is caused by a decision or action made by humans
	AI	The risk is caused by a decision or action made by an AI system
	Other	The risk is caused by some other reason or is ambiguous
Intent	Intentional	The risk occurs due to an expected outcome from pursuing a goal
	Unintentional	The risk occurs due to an unexpected outcome from pursuing a goal
	Other	The risk is presented as occurring without clearly specifying the intentionality
Timing	Pre-deployment	The risk occurs before the AI is deployed
	Post-deployment	The risk occurs after the AI model has been trained and deployed
	Other	The risk is presented without a clearly specified time of occurrence

[28] Peter Slatery et al, The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence, an extensive 80 pages document. Published on above link, 5.

Table B. Domain Taxonomy of AI Risks

Domain / Subdomain	Description
1 *Discrimination & toxicity*	
1.1 Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.
1.2 Exposure to toxic content	AI that exposes users to harmful, abusive, unsafe, or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
1.3 Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions are dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.
2 *Privacy & security*	
2.1 Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information	AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectation of privacy, assist identity theft, or cause loss of confidential intellectual property.
2.2 AI system security vulnerabilities and attacks	Vulnerabilities that can be exploited in AI systems, software development toolchains, and hardware that results in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.
3 *Misinformation*	
3.1 False or misleading information	AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans that make decisions based on false beliefs can experience physical, emotional, or material harms
3.2 Pollution of information ecosystem and loss of consensus reality	Highly personalized AI-generated misinformation that creates "filter bubbles" where individuals only see what matches their existing beliefs, undermining shared reality and weakening social cohesion and political processes.
4 *Malicious actors & misuse*	
4.1 Disinformation, surveillance, and influence at scale	Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim of manipulating political processes, public opinion, and behavior.
4.2 Cyberattacks, weapon development or use, and mass harm	Using AI systems to develop cyber weapons (e.g., by coding cheaper, more effective malware), develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or chemical, biological, radiological, nuclear, and high-yield explosives), or use weapons to cause mass harm.
4.3 Fraud, scams, and targeted manipulation	Using AI systems to gain a personal advantage over others through cheating, fraud, scams, blackmail, or targeted manipulation of beliefs or behavior. Examples include AI-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.
5 *Human-computer interaction*	
5.1 Overreliance and unsafe use	Anthropomorphizing, trusting, or relying on AI systems by users, leading to emotional or material dependence and to inappropriate relationships with or expectations of AI systems. Trust can be exploited by malicious actors (e.g., to harvest information or enable manipulation), or result in harm from inappropriate use of AI in critical situations (such as a medical emergency). Overreliance on AI systems can compromise autonomy and weaken social ties.

Table B. Domain Taxonomy of AI Risks

Domain / Subdomain	Description
1 *Discrimination & toxicity*	
1.1 Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.
1.2 Exposure to toxic content	AI that exposes users to harmful, abusive, unsafe, or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
1.3 Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions are dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.
2 *Privacy & security*	
2.1 Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information	AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectation of privacy, assist identity theft, or cause loss of confidential intellectual property.
2.2 AI system security vulnerabilities and attacks	Vulnerabilities that can be exploited in AI systems, software development toolchains, and hardware that results in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.
3 *Misinformation*	
3.1 False or misleading information	AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans that make decisions based on false beliefs can experience physical, emotional, or material harms
3.2 Pollution of information ecosystem and loss of consensus reality	Highly personalized AI-generated misinformation that creates "filter bubbles" where individuals only see what matches their existing beliefs, undermining shared reality and weakening social cohesion and political processes.
4 *Malicious actors & misuse*	
4.1 Disinformation, surveillance, and influence at scale	Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim of manipulating political processes, public opinion, and behavior.
4.2 Cyberattacks, weapon development or use, and mass harm	Using AI systems to develop cyber weapons (e.g., by coding cheaper, more effective malware), develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or chemical, biological, radiological, nuclear, and high-yield explosives), or use weapons to cause mass harm.
4.3 Fraud, scams, and targeted manipulation	Using AI systems to gain a personal advantage over others through cheating, fraud, scams, blackmail, or targeted manipulation of beliefs or behavior. Examples include AI-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.
5 *Human-computer interaction*	
5.1 Overreliance and unsafe use	Anthropomorphizing, trusting, or relying on AI systems by users, leading to emotional or material dependence and to inappropriate relationships with or expectations of AI systems. Trust can be exploited by malicious actors (e.g., to harvest information or enable manipulation), or result in harm from inappropriate use of AI in critical situations (such as a medical emergency). Overreliance on AI systems can compromise autonomy and weaken social ties.

The detailed **table B**[29] then shows the Domain taxonomy of AI Risks (ethical challenges). The first four domains are the most discussed: discrimination biases and toxicity, privacy and security, misinformation, and malicious actors. Less discussed but ethically as important are the domains 5-7 in the table: Human-computer interaction, socioeconomic and environmental harms, AI system safety, failures, and limitations.

These taxonomies A and B are impressively detailed with the 700 risk parameters. Nevertheless, the economic and financial risks on a company level, industry level, banking level and macroeconomic level are in my view underestimated and only partly reflected in B6 and B7.

This taxonomy and my ethical reading of it is misunderstood if it gives the impression, ethics looks only at risks of AI and does not see the benefits and changes. On the contrary: AI offers many opportunities and potentials to contribute solving problems of humanity. All 700 risk parameters could be in fact translated into positive opportunities.

1.5 Ethics in the co-driver seat, not on the emergency brake: Seven recommendations

Towards the end of this article, I come back to the questions at the beginning about the role of ethics in AI governance and regulations: Are ethicists on the emergency brake of the car? Are they in the co-driver seat? Are they on the back seat or in the developers' lab? What are the phases in technological development where ethics comes in? Let me conclude with some recommendations.

1. *Every person can be and can become a values-driven actor*. Values-driven and not only power-driven or money-driven is a quality of humans independent from status, education, gender, culture and religion.

[29] Ibid, 6-7.

Especially leaders wherever they are responsible for AI-related topics, need to be and can be or become values-driven actors.

2. *Specialised, trained ethicists are needed*. In addition to the call to everybody to decide and act based on their values, ethics specialists are needed. A key training is to deal with ethical dilemmas – the conflict between competing values or anti-values. My observation during last decades is that ethics became almost mainstream as a requirement in business, but often in the hands of not enough trained specialists. An IT specialist, a manager or a lawyer who followed a one semester voluntary or compulsory crash course on ethics is good, but by far not enough! I speak as a professor of ethics who trained students on all continents. Profound philosophical, theological, cultural studies on value-systems and their implications are needed at the developers' and the implementers' desks.

3. *Many ethics officers need more profound ethics background.* They often are not enough trained in ethics. The global ethics and compliance industry, especially in its form in the Anglo-Saxon world, is often leading to mainly legal compliance department, which checks the boxes of legal requirements. This is far from enough for an ethics department. AI ethics is a matter of the Board room and top management with strategic ethics competences!

4. Ethics must be in the co-driver seat and not only on the emergency brake! The history of technology shows pattern: New technological inventions and innovations lead to enthusiasm and a "gold rush" atmosphere, almost irrational unregulated rush not to lose the opportunity. Then the first accidents happen, the abrupt crash or stop with the emergency brake happens. The early warners are then confirmed but could not hinder the crash. Emergency helpers like tech-doctors, ethicists, regulators are then called to repair and advise how to avoid further crashes. This behaviour causes limited damage in the stone age, when an early

harpoon kills a human being by accident instead of the bear or a test car is demolished. But AI immediately has an influence on a global scale. This is why even innovators and most powerful entrepreneurs like Elon Musk asked in an early phase of the current boom for a six-month moratorium on large-scale AI development to allow time to properly assess the risks that might arise, and then for AI regulations and effective regulators. The other model is, to put ethics in the co-driver seat. Ethicists together with technologists, sociologists, regulators, educators, and investors develop in a multidisciplinary team the new technologies. This already happens, but often not in an early stage, only when the technology is almost ready for the market. Two examples: I served for ten years at the beginning of this century in the advisory commission of the Swiss government for bioethics in the non-human field (Swiss Federal Commission on Bioethics in the Non-Human Field EKAH, about Genetically modified crops, animal testing etc.). We advised the Swiss government e.g., if a genetically modified wheat can go to be implemented in the agricultural field. The industry refused to have the ethicists at the table during the initial design phase of the GMO-products. The opposite example is CERN in Geneva: I attended on 3 March 2024 a workshop of its new Open Quantum Institute. I was impressed how CERN underlined that ethics need to be at the table from the very beginning of the design of new technologies and therefore in this new institute. That is what I mean by Co-driver seat. It is not about power, it is about including ethics in the very early phase of development, where the ethical questions are not yet clear. The two examples also show the difference between private sector with high IP protection and the public sector such as CERN with its pro-active open access policy.

5. *Ethical reflections and strong cooperation between the private and public sector are needed*. AI is cross-sectoral of high relevance and therefore needs trustworthy cooperation between private, public, academic,

and civil society sectors. This simple sentence has huge geopolitical systemic implications: in the US, the private AI giants want to be as independent from government regulations and influence. In China, public AI regulations have been among the first in the world and private sector from a very early phase must cooperate with the government. Behind is the difference between market and plan economy. Europe with its middle way wants to allow as much as possible the free market, but also develops strong regulations as with the new EU AI Act of 2024. The European socially and environmentally regulated market economy is the model between US and China.

6. *Multilateral standardisation and regulation organisations need to be strengthened*. The World Intellectual Property Organisation ("WIPO") and the International Telecommunication Union ("ITU"), both with their headquarter in Geneva, together with UNESCO in Paris with its educational mandate are three of the important regulatory bodies for AI governance. This shows how important multilateral organisations in the UN system are, in cooperation with voluntary standard organisations such as ISO or IEEE.

7. *Ethical investment criteria* need to be applied also in AI and related new technological developments. The investment sector made huge progress in the last four decades in emphasising and trying to mainstream ethical and environmental criteria in investments. Responsible investments, ethical investments and sustainable investing became mainstream terms. Related specialised and qualified rating organisations have been built. The process of including AI Ethics in ethical investing is at the beginning and needs to be further unfolded.

In conclusion: The fascinating, very fast development of Artificial Intelligence, soon combined with Quantum computing as the next development leap, opens many opportunities for humanity and includes

manifold risks. To reduce risks and increase positive impact on human and environmental development is the reason, why AI Governance Ethics is urgently and timely need.

2

AI DIPLOMACY: NEGOTIATING INTERNATIONAL AI GOVERNANCE AND REGULATION

Jovan Kurbalija, Switzerland / Serbia

Artificial intelligence[30] (AI) has rapidly evolved from a niche technological innovation to a transformative force reshaping industries, economies, and societies worldwide. As AI systems become increasingly integrated into critical sectors such as healthcare, finance, transportation, and national security, the need for robust governance and regulation has never been more urgent.

The international community faces the complex challenge of balancing innovation with ethical considerations, security concerns, and socio-economic impacts. This analysis delves into the role of diplomacy in negotiating AI governance and regulation, by exploring a few inquiry

[30] *Jovan Kurbalija* is the Founding Director of DiploFoundation and the Head of the Geneva Internet Platform. A former diplomat, he has a professional and academic background in international law, diplomacy, and information technology. He has been a pioneer in the field of cyber diplomacy since 1992 when he established the Unit for Information Technology and Diplomacy at the Mediterranean Academy of Diplomatic Studies in Malta. This text is an excerpt from his forthcoming publication on 'AI Diplomacy'.

questions: Why is diplomacy important for AI governance? Who are the main actors? How is AI negotiated? Where are negotiations happening? What are the main AI governance issues?

2.1 Why is diplomacy important for AI governance?

Diplomacy, the art, and practice of conducting negotiations between representatives of states, has historically played a crucial role in addressing global challenges. Diplomacy takes on a new role in negotiations on the governance of AI for several reasons.

First, AI technologies often transcend national borders, necessitating international cooperation to address data, security, and economic policy issues. Second, the rapid pace of AI development requires agile and adaptive international regulatory frameworks that can keep up with technological advancements. Third, the geopolitical implications of AI, including its potential to shift power dynamics and trigger conflicts, underscore the need for coordinated international AI governance.

2.2 Who are the main actors in AI governance?

AI governance is a policy space in the making. Various actors are finding their way in the new stratification of power and roles. There are three main types of actors in AI governance: national governments and supra-national organisations like the EU, tech companies and business associations, and international organisations. We tackle the first two categories below and discuss international organisations in the section dedicated to *WHERE* is AI diplomacy performed?

2.2.1 Governments

Governments are searching for the right approach to AI governance and policy, reflecting the fast-changing nature of AI developments. Three

actors are taking the lead: the USA and China are AI superpowers, and the EU is the leader in regulatory issues. Japan and India are also very active actors in the field of AI diplomacy.

United States

The United States has been at the forefront of AI developments. The country has published multiple AI strategies and policies, emphasising, among other issues, the need to maintain American leadership in AI. Various sectoral AI plans and initiatives in areas such as defence, air force, energy, standardisation, and research and development have been published over the years. This multi-faceted approach highlights the USA's comprehensive strategy to integrate AI across various sectors, ensuring its leadership in both civilian and military applications.

On the international level, the USA reaffirmed its commitment to AI leadership by participating in various international initiatives and promoting the development of AI standards through organisations like the Institute of Electrical and Electronics Engineers (IEEE) and the International Organization for Standardization (ISO). The USA has also been involved in discussions on the implications of AI in areas such as ethics, human rights, and sustainable development; for instance, it was an active participant in the negotiations on the Council of Europe’s convention on AI and human rights and was behind the UN General Assembly’s first resolution on AI.

China

China has set ambitious goals to become a world leader in AI by 2030, as outlined in its "New Generation Artificial Intelligence Development Plan". The country has invested heavily in AI research, development, and infrastructure, including establishing a $2.1 billion technology park dedicated to AI companies. China's strategy emphasises the integration of AI

into various sectors, from healthcare to autonomous vehicles, and seeks to leverage AI for economic and military advancements.

China is active internationally in standardisation and policy bodies. China's priority is to have AI governance at the UN as indicated by the call of President XI[31] to 'strengthen the governance of artificial-intelligence rules within the framework of the United Nations'. At the UN General Assembly level, China was behind another resolution on AI and capacity development (adopted a few months after the US-proposed resolution we mentioned above). China also proposed a framework for global AI governance[32] – under the name Global AI Governance Initiative – calling on countries to 'work together to prevent risks, and develop AI governance frameworks, norms and standards based on broad consensus, so as to make AI technologies more secure, reliable, controllable, and equitable.'

European Union

The European Union (EU) has taken a comprehensive approach to AI governance, focusing on ethical considerations and regulatory frameworks. The EU's AI Act aims to govern AI largely at the level of applications/uses, ensuring that the technology is transparent, reliable, and safe. The EU also emphasises the importance of data governance, advocating for the digitalization and sharing of publicly held data to enhance

---

[31] https://www.economist.com/china/2024/08/25/is-xi-jinping-an-ai-doomer?Utm_campaign=r.the-economist-sunday-today&utm_medium=email.internal-newsletter.np&utm_source=salesforce-marketing-cloud&utm_term=20240825&utm_content=ed-picks-image-link-3&etear=nl_sunday_today_3&utm_campaign=r.the-economist-sunday-today&utm_medium=email.internal-newsletter.np&utm_source=salesforce-marketing-cloud&utm_term=8/25/2024&utm_id=1915102

[32] https://dig.watch/updates/china-launches-global-ai-governance-initiative

AI capabilities. The EU's stance on AI governance reflects its broader commitment to digital sovereignty and ethical standards.

The main regulatory instrument is the **EU AI Act** which provides a risk-based regulatory approach for AI systems.

Japan

A commitment to ethical AI development, international cooperation, and the harmonisation of technical standards characterises Japan's diplomatic stance on international AI governance negotiations. The country actively participates in global forums such as the G7, OECD, and Internet Governance Forum, advocating for responsible AI practices that align with universal human rights and ethical principles. Japan's strategic objectives include maintaining technological leadership, ensuring data sovereignty, and promoting human-centred AI. As AI continues to evolve, Japan's balanced approach to innovation and regulation will be crucial in shaping the future of global AI governance.

Japan recognises the importance of data as a strategic asset in AI development. The country advocates for the digitalization of publicly held data to enhance AI-related capacities. Japan has been an active participant in the G7 and OECD discussions on AI governance. The Hiroshima G7 Leaders' Communiqué underscores Japan's commitment to advancing multi-stakeholder approaches to AI standard development, emphasising transparency, openness, fairness, privacy, and inclusiveness (Hiroshima G7 Leaders' Communiqué)[33]. Japan's involvement in the G7 working group on generative AI, established through the Hiroshima AI process, further demonstrates its proactive role in shaping global AI governance frameworks (Hiroshima G7 Leaders' Communiqué).

[33] Hiroshima G7 Leaders' Communiqué. Retrieved from https://carnegieendowment.org

India

A proactive and inclusive approach characterises India's international AI governance negotiations diplomacy. The country has leveraged its national AI strategy, multilateral diplomacy, and commitment to ethical AI to shape global AI policy frameworks. By advocating for inclusive representation, promoting AI for sustainable development, and supporting international standards, India aims to contribute to a coherent and equitable global AI ecosystem. As AI continues to evolve, India's strategic position and collaborative efforts will play a crucial role in shaping the future of international AI governance.

2.2.2 Tech companies

The AI and tech industry is represented in international negotiations directly or through various business associations and lobbyist groups. They have diverse positions, as indicated by California's proposed AI regulation, which is supported by the company Anthropic while opposed by many other AI giants.

Here is a survey of some of the main actors involved in AI governance and regulation discussions and negotiations.

OpenAI

OpenAI's position on AI governance and regulation emphasises the importance of regulating AI systems, especially those with significant capabilities, to ensure safety, security, and ethical use. OpenAI acknowledges the necessity of regulation while cautioning against both excessive regulation and inadequate oversight.

The company is actively engaging with policymakers globally to help shape regulations that balance safety with accessibility to AI technology's benefits. However, there are major gaps where calls for regulation start being practical and operational. A recent example is the company’s

opposition to the California's AI regulation[34] (SB 1047). This gap between rhetoric and actions should be looked at carefully by policymakers, especially when they consider OpenAI's statements on a number of AI policy issues that they have been arguing for, including transparency, democratic oversight, dealing with misuses, etc.

Microsoft

Microsoft's priorities in the AI field are closely aligned with OpenAI, which Microsoft strongly supports financially and in regulatory lobbying. At the same time, Microsoft tries to keep some distance from OpenAI in order not to get any negative regulatory spillover from AI regulations targeted largely at OpenAI.

Microsoft is actively advocating for implementing guidelines and laws to govern the use of AI technologies. The company has released a 40-page report titled 'Governing AI: A Blueprint for the Future'[35], highlighting the importance of increased regulation in the AI industry. Microsoft President Brad Smith emphasised the need for shared responsibility in establishing the necessary guardrails for AI systems. By promoting AI regulations, Microsoft aims to create a framework that fosters transparency, accountability, and fairness in AI systems while ensuring that AI technologies align with human values and have a positive societal impact.

Microsoft acknowledges the potential risks associated with AI, including privacy concerns, bias in algorithms, and the potential displacement of jobs. The company believes that regulations can help address these issues and facilitate the ethical and responsible development and deployment of AI technologies.

---

[34] https://dig.watch/updates/openai-opposes-californias-ai-regulation-bill

[35] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

Google (Alphabet Inc.)

Google's position on AI governance and regulation is centred around its "Opportunity Agenda"[36] for AI regulation. Google emphasises the importance of a global approach to regulating AI to prevent a fragmented regulatory landscape that hinders innovation and creates barriers for smaller companies. The company advocates for transparency and public input in the decision-making process for AI regulation, aiming to build public trust and ensure that regulations reflect societal values and concerns.

The "Opportunity Agenda" focuses on four key pillars: safety, fairness, privacy, and accountability. Google aims to address these aspects to ensure that AI technologies are not only cutting-edge but also adhere to ethical standards. The company calls for governments to set clear guidelines and standards for global AI development and deployment, emphasising collaboration between industry leaders, academic institutions, and civil society to shape a harmonised global approach to AI regulation.

Anthropic

Anthropic promotes AI governance centralised around multi-stakeholder engagement, international collaboration, and the incorporation of human rights standards in AI development.

Typically, Anthropic aligns with the position of other tech companies. However, sometimes it takes separate approach, as was the case with the California AI regulation.

Facebook (Meta)

Meta's position on AI governance is shaped by the company's strategic focus on open-source AI built around the Llama platform. Such an approach makes a major difference with OpenAI and other companies,

[36] https://dig.watch/updates/googles-opportunity-agenda-global-approach-to-responsible-ai-regulation

which argue that closed AI platforms are needed to deal with security and other risks of the open-source approach.

Facebook is also for international governance, and it should be built around an international agency for AI governance.[37] The company emphasises responsible AI development by following principles prioritising security, privacy, accountability, transparency, diversity, and robustness. Meta's commitment to transparency includes open sourcing its large language model to allow external scrutiny and collaboration within the AI community.

Nvidia

Nvidia is the main manufacturer of GPUs (Graphical Processing Units), which are used to develop foundational modules. The more AI development is decentralised, the more GPUs they can sell.

Thus, it is not surprising that Nvidia strongly supports countries building their own AI infrastructure to preserve cultural identities and leverage the economic benefits of AI. The main policy concern of Nvidia is compliance with export restrictions imposed by the US. Nvidia's products are often targeted by export restrictions, especially towards China.

2.2.3 Business associations and lobbyists

These actors are involved in lobbying, conducting policy research, organising events, and providing platforms for collaboration among AI companies and stakeholders. They aim to influence policy decisions, promote industry interests, and shape public perceptions of AI technologies.

The Partnership on AI (PAI) is one of the most prominent associations representing the AI industry. Founded by major tech companies, including Amazon, Apple, DeepMind, Facebook, Google, IBM, and Microsoft, PAI aims to address the opportunities and risks associated with AI

---

37 https://dig.watch/updates/metas-global-policy-chief-warns-governments-against-fragmented-laws-around-ai

(The Rise of TechPlomacy in the Bay Area).[38] The organisation focuses on researching AI and tackling its scientific and societal impacts. PAI's initiatives include projects like Microsoft's AI for Earth and Facebook's use of AI to combat online terrorism.

Other business associations involved in AI governance include:

- **TechNet:** A network of technology CEOs and senior executives that advocates for the growth of innovation economy, including AI.
- **Business Software Alliance (BSA)** is one of the most established digital industry associations with a long tradition in lobbying policymakers.
- **Computer & Communications Industry Association (CCIA)** is a business association promoting open systems, and open networks in the tech sector, including AI.
- **World Economic Forum (WEF)** brings together leading companies in AI to shape the global AI agenda through the AI Forum.

2.2.4. International organisations

International organisations provide venues for negotiations on AI governance, as indicated in the section 'HOW" of this text. However, they also play in AI governance negotiations and processes. ITU, UNESCO, OECD, the Council of Europe, G7, and G20 are the most active players.

2.3 Where is AI diplomacy performed?

Here is a survey of the main AI development and governance hubs:

Bay Area, USA is the epicentre of AI developments, with a high concentration of talent, investment, and innovation. In San Francisco and

[38] https://dig.watch/trends/artificial-intelligence

Silicon Valley, there are the headquarters of tech giants like Google, Meta, and OpenAI that shape AI technologies and trends. The critical relevance of the Bay Area's AI developments has attracted many countries to establish, so-called, tech diplomacy representation in the Bay Area.

Washington D.C., USA as the political heart of the United States, plays a pivotal role in shaping AI governance. AI is high on the agenda of the U.S. Congress, the National Telecommunications and Information Administration (NTIA), and other key federal agencies that are instrumental in crafting AI policies. The Biden administration has already taken steps towards new AI rules, studying potential accountability measures for AI systems like ChatGPT. The city is home to various federal agencies, think tanks, and academic institutions that can collaborate to address the multi-disciplinary challenges posed by AI.

AI diplomacy initiatives are coordinated by the U.S. State Department, which established the Bureau of Cyberspace and Digital Diplomacy and the Office of the Special Envoy for Critical and Emerging Technology.

New York, USA is the seat of the United Nations (UN) which coordinates international initiatives in AI governance. The UN's involvement is multifaceted, encompassing various specialized agencies and initiatives to cover various sectoral aspects of AI, from culture to development and security.

The holistic policy approach to AI is conducted in the AI Advisory Group, which produced the UN's Interim Report: Governing AI for Humanity[39] underscores the importance of a universal, networked approach to AI governance, rooted in the public interest and inclusive of all stakeholders. Here are other specific coverages of AI:

[39] https://www.un.org/en/content/digital-cooperation-roadmap/

Brussels, Belgium, hosts the European Union institutions and regulatory capital of Europe with a high impact on digitalisation and AI. The so-called 'Brussels effect' has a far-reaching impact on digital and AI governance worldwide. The General Data Protection Regulation (GDPR), which came into effect in 2018, is a prime example. GDPR has not only set the standard for data protection within the European Union (EU) but has also influenced data privacy laws worldwide, including in countries like Brazil, Japan, and even China (Source: AI and Digital in 2023: From a winter of excitement to an autumn of clarity).

It is expected that the **EU AI Act**, which came into force on 1st August 2024, will have a considerable regulatory impact on AI developments. Brussels, as a regulatory powerhouse, has attracted numerous international organisations, permanent missions, think tanks, and non-governmental organisations (NGOs) that play a crucial role in shaping global digital policies.

Beijing, China, has emerged as an AI hub reflecting China's overall AI developments and, in particular, AI regulations and ethical guidelines. Beijing is the seat of many prominent AI companies and governance authorities, including the Cyberspace Administration of China (CAC) which has been developing AI governance norms. The China Academy of Information and Communications Technology (CAICT) has been working on developing testing and certification systems for 'trustworthy AI'. Beijing is also the host of the Global AI Governance Initiative (GAIGI), which aims to bring together 155 countries participating in the Belt and Road Initiative.

Geneva, Switzerland, shapes its role as an AI governance hub on the long tradition of governing technological developments for more than a century since the establishment of the **International Telecommunication Union**. Geneva hosts the annual AI for Good Global Summit as the main event linking AI and development interplay. The city also hosts

'trinity' of the main digital standardisation organisations: International Telecommunication Union, International Organization for Standardization, and International Electrotechnical Commission. Geneva hosts a wide range of specialised UN agencies that use AI to reform their activities in the field of health (World Health Organisation), trade (World Trade Organisation and UNCTAD), environment (World Meteorological Organisation), and humanitarian (International Committee on Red Cross).

Paris, France is home to several key international organisations that play significant roles in AI governance. The Organisation for Economic Co-operation and Development (OECD), headquartered in Paris, has been instrumental in developing AI principles that emphasise human rights, transparency, and accountability. These principles have been adopted by numerous countries, setting a global standard for AI ethics.

The United Nations Educational, Scientific and Cultural Organization (**UNESCO**), also based in Paris, has adopted a global agreement on AI ethics. Paris is also the seat of the "Paris Call for Trust and Security in Cyberspace", launched in 2018 by President Macron. The hosting of OECD and UNESCO, initiatives such as Paris Call, and France's prominent role in shaping the EU AI Act make France and Paris an important hub of AI governance.

London, UK, is a burgeoning AI hub with many businesses and research institutions. UK's ambitions to become a global leader in AI governance were outlined in "The UK's International Technology Strategy" (2023). The implementation started with the hosting of the AI Safety Summit in November 2023 and the adoption of the Bletchley Declaration on AI Safety.[40]

[40] https://dig.watch/resource/the-bletchley-declaration

2.4 How is AI diplomacy performed? Manifold actors

AI diplomacy is in search of *modus operandi.* Mechanisms and approaches are constantly designed and adjusted. We start with a mapping of governance mechanisms that exist in the private sector and move to the national level, which shapes the interests that diplomatic services must protect. Lastly, we discuss mechanisms on an international level.

2.4.1 Private sector self-regulation

The private sector is instrumental in pushing the technological development of AI. In addition to investing in research and development activities, companies in the AI field also pay attention to the need to acquire the trust of their customers. This is one of the reasons why companies have—over the years—expressed their commitments to safe, secure, and trustworthy AI (in these or similar words) through a wide range of principles, guidelines, and codes of conduct.

Google, for instance, has a set of AI Principles[41] – periodically updated – and a Secure AI Framework[42] that are meant to guide the company's responsible development of AI technology and to address concerns related to issues such as AI risk management, security, and privacy.

Microsoft's Responsible AI Principles[43] highlight issues such as fairness, reliability and safety, privacy and security, and accountability as being at the core of the company's approach to designing, building, and testing AI systems. The company has also developed a Responsible AI Standard[44] to operationalise these principles. Tencent also has a set of AI Principles according to which AI should be available, reliable,

[41] https://ai.google/responsibility/principles/

[42] https://safety.google/cybersecurity-advancements/saif/

[43] https://www.microsoft.com/en-us/ai/principles-and-approach

[44] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl?culture=en-us&country=us

comprehensible, and controllable. IBM's Principles for Trust and Transparency[45] highlight three main commitments: the purpose of AI is to augment human intelligence, data and insights belong to their creators, and AI must be transparent and explainable.

In addition to these and similar principles and guidelines released by individual companies, there are also instances of several companies coming together to highlight their commitments related to trustworthy AI. For instance, in July 2023, Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI signed up to a set of voluntary commitments[46] to advance safe, secure, and transparent development and use of AI technology. In September 2023, eight other companies – Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability – signed up to an additional set of commitments[47] particularly related to generative AI model technology. Another example comes from the Frontier Model Forum,[48] an initiative launched in July 2023 by Anthropic, Google, Microsoft, and OpenAI, and dedicated, among other issues, to advancing work on safety standards and evaluations to ensure frontier AI models are developed and deployed responsibly.

---

45 https://www.ibm.com/policy/trust- principles/#%3Atext%3DWe%20believe%20that%20government%20data%2Cand%20equitable%20and%20prioritize%20openness.%26text%3DClients%20are%20not%20required%20to%2Cof%20IBM%27s%20solutions%20and%20services.%26text%3DIBM%20client%20agreements%20are%20transparent

46 https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

47 https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf

48 https://www.frontiermodelforum.org/

2.4.2 National governance approaches

Soft law approaches

Some governments have decided to take more of a hands-off approach towards AI and are not – at least for the time being – moving into developing specific regulations. They argue, on the one hand, that existing regulations in areas such as product safety, competition, and data protection are equally applicable to AI, and, on the other, that the implications of AI for the economy and society are yet to be fully understood, and as such regulation might be premature.

Switzerland, for instance, has so far taken a so-called 'technological neutrality' approach, arguing that existing laws and regulations are sufficient to address AI-related challenges. While these laws and regulations may require some updates to reflect AI developments, there is no need to create new, AI-specific instruments (Swiss Federal Department of Foreign Affairs, 2022).

Singapore is another example worth noting. The country has taken a soft law approach by providing guidance on how organisations could adopt and deploy AI responsibly with the release of a Model AI Governance Framework[49] (2020, 2nd edition). This is complemented with an AI governance testing framework and toolkit – A.I. Verify[50] – to enable the industry to demonstrate their deployment of responsible AI. Singapore also has a Model AI Governance Framework for Generative AI,[51] outlining principles and practical suggestions to support a comprehensive and trusted AI ecosystem, as well as a Generative AI Evaluation Sandbox[52] to enable the evaluation of trusted generative AI products.

---

[49] http://go.gov.sg/ai-gov-mf-2

[50] https://file.go.gov.sg/aiverify.pdf

[51] https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf

[52] https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

Regulations

Regulation is one type of governance that is led by states. In general, such regulation can take two forms: horizontal, covering AI systems/applications broadly, and sectoral, targeted at specific types of AI applications.

One of the most notable examples of horizontal regulation is the EU AI Act[53]. Proposed in spring 2021 and approved in June 2024, the Act introduces a risk-based regulatory approach for AI systems:

- **If an AI system poses exceptional risks, it is banned.** This applies, for instance, to AI systems that manipulate human behaviour to circumvent free will; biometric categorisation systems that use sensitive characteristics (e.g., political, religious, philosophical beliefs, sexual orientation, race); social scoring based on social behaviour or personal characteristics; untargeted scraping of facial images from the internet or CCTV footage to create facial recognition databases; and emotion recognition in the workplace and educational institutions.
- **If an AI system comes with high risks** (for instance, the use of AI in performing surgeries), it will be **strictly regulated**. Such systems would need to be closely assessed before being put on the market, and they would need to comply with a set of requirements such as having a risk management system in place; having appropriate data governance and management practices for training, validation, and testing data sets; enabling automatic record-keeping of events; being effectively overseen by humans; and achieving an appropriate level of accuracy, robustness, and cybersecurity. There are also certain instances under which the deployers of high-risk AI systems (such as public institutions) are required to perform assessments of the impact on fundamental rights that the use of the system may produce.

[53] https://eur-lex.europa.eu/eli/reg/2024/1689/oj

- **If an AI system only involves limited risks, focus is placed on ensuring transparency for end users.** Unlike horizontal regulations, which cover a wide range of AI systems and applications, sectoral or specific regulations target certain types of AI applications or certain AI uses.

China has so far taken such an approach to the regulation of AI systems, and there are three key AI regulations the country has adopted in recent years: the 'Regulations on the Management of Algorithm Recommendation for Internet Information Services' [54](2021), the 'Regulations on the Management of Deep Synthesis Internet Information Services'[55] (2022) , and the 'Interim Measures for the Management of Generative Artificial Intelligence Services'[56] (2023).

Around the world, other jurisdictions have also passed sector-specific regulation (at either the national or local level) to cover certain uses or applications of AI, such as autonomous vehicles and their testing/use on public roads, and the use of facial recognition technology by law enforcement agencies.

In between soft and hard law?

The USA has traditionally taken more of a soft-law approach towards the governance of technology, and this is also the case with AI (at least so far). In 2022, the White House – through the Office of Science and Technology – released a Blueprint for an AI Bill of Rights[57] outlining a

---

[54] https://clairk.digitalpolicyalert.org/documents/china-regulations-on-the-management-of-algorithm-recommendation-for-internet-information-services/

[55] https://clairk.digitalpolicyalert.org/documents/china-regulations-on-the-management-of-deep-synthesis-internet-information-services/

[56] https://clairk.digitalpolicyalert.org/documents/china-interim-measures-for-the-management-of-generative-artificial-intelligence-services/

[57] https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

set of principles to guide the design, use, and deployment of automated systems to protect citizens.

The following year, in October 2023, President Biden signed the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[58]. Among the multitude of provisions, the order places a heavy emphasis on AI safety and security issues. Most of this emphasis takes a rather soft law approach. For instance, the order requires several federal agencies to work together and develop guidelines to enable AI developers to conduct AI red-teaming tests to enable the deployment of safe, secure, and trustworthy AI systems.

A more hard law approach is taken with regard to dual-use foundation models: here, the Secretary of Commerce is to require companies that develop or demonstrate an interest to developing potential dual-use foundation models to comply with certain reporting requirements (e.g. the cybersecurity measures taken to ensure the integrity of the training process, the measures taken to protect model weights, and the results of any dual-use foundation model's performance in relevant AI red-team testing).

2.4.3 Multilateral initiatives

AI has been on the agenda of UN agencies and several other *minilateral* and multilateral organisations (be they regional or global) for several years now.

Some of these agencies and organisations have been exploring how AI is or may be impacting their areas of work, therefore following the sectoral approach mentioned above. UNICEF, for instance, has been working on policy guidance[59] to promote children's rights in government and private sector AI policies and practices, while the International

[58] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[59] https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

Labour Organization (ILO) is looking at the impact of AI (including generative AI)[60] and automation on the world of work.

The World Intellectual Property Organization (WIPO) is discussing intellectual property issues[61] related to the development of AI, whereas the World Health Organization (WHO) looks at the applications and implications of AI in healthcare, and related governance[62] approaches. ITU is hosting an annual AI for Good summit[63] exploring the use of AI to accelerate progress towards sustainable development, in addition to having several study groups and focus groups carrying out AI-related standardisation and pre-standardisation work.

Other entities have focused on developing horizontal frameworks for AI. Three key examples include:

- The Recommendation on AI[64] adopted in May 2019 by the Organization for Economic Co-operation and Development (OECD).

The recommendation outlines a set of principles for the responsible stewardship of trustworthy AI and calls on AI actors to work towards AI that promotes inclusive growth, sustainable development, and well-being, and to implement mechanisms and standards that support the design of AI systems in line with the rule of law, human rights, democratic values, and diversity.

- The Recommendation on the ethics of AI adopted by the United Nations Educational, Scientific and Cultural Organisation[65] (UNESCO)

---

60 https://www.ilo.org/publications/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and

61 https://www.wipo.int/about-ip/en/frontier_technologies/ai_and_ip.html

62 https://www.who.int/news/item/19-10-2023-who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health

63 https://www.itu.int/en/action/ai/Pages/default.aspx

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

65 https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

in 2021. The recommendation outlines a series of values, principles, and actions to guide states in the formulation of their legislation, policies, and other instruments regarding AI. For instance, the document calls for action to guarantee individuals more privacy and data protection, by ensuring transparency, agency, and control over personal data. Explicit bans on the use of AI systems for social scoring and mass surveillance are also highlighted, and there are provisions for ensuring that real-world biases are not replicated online.

- The Framework Convention on AI and Human Rights, Democracy and the Rule of Law[66] adopted by the Council of Europe in 2024. The convention outlines key principles to be complied with throughout the lifecycle of AI systems: human dignity and individual autonomy, equality and non-discrimination, respect for privacy and personal data protection, transparency and oversight, accountability and responsibility, reliability, and safe innovation.

Within the UN system, in addition to the work carried out by various UN agencies in relation to AI, the technology and its policy implications have also attracted the attention of the UN Secretary-General[67], the UN General Assembly (UNGA), and even the Security Council[68]. In 2024, UNGA adopted two resolutions on AI:

---

[66] https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[67] https://www.un.org/sg/en/content/sg/speeches/2023-10-26/secretary-general%E2%80%99s-remarks-press-conference-launching-high-level-advisory-body-artificial-intelligence%C2%A0#%3A~%3Atext%3DFor%20developing%20economies%2C%20AI%20offers%2Cis%20difficult%20even%20to%20grasp

[68] https://dig.watch/event/the-uk-is-angling-for-global-regulation-of-ai-through-the-un-security-council

- Resolution 78/265[69]: Seizing the opportunities of safe, secure, and trustworthy artificial intelligence systems for sustainable development (proposed by the USA); and
- Resolution 78/311[70]: Enhancing international cooperation on capacity-building of artificial intelligence (proposed by China).

AI governance is also tackled in the Global Digital Compact[71] (GDC), approved late September by UN member states in the context of the Summit of the Future in September 2024. It is worth noting that discussions around AI governance have revolved around three main proposals (and variations of them): an International Scientific Panel on AI, an annual Global Dialogue on AI Governance, and a Global Fund for AI for Sustainable Development.

Some of these proposals are inspired by recommendations from the multistakeholder High-level Advisory Body on AI[72] established by the UN Secretary-General in 2023 to 'advance recommendations for the international governance of AI'.

The following specialised organisations take leading role in AI governance processes:

The **UN Educational, Scientific and Cultural Organization** (**UNESCO**) has been at the forefront of developing ethical guidelines for AI. In 2021, UNESCO adopted the Recommendation on the Ethics of Artificial Intelligence, calling member states to establish legal and institutional frameworks. to govern AI technologies. This recommendation

---

[69] https://undocs.org/Home/Mobile?FinalSymbol=A%2FRES%2F78%2F265&Language=E&DeviceType=Desktop&LangRequested=False

[70] https://www.undocs.org/Home/Mobile?FinalSymbol=A%2FRES%2F78%2F311&Language=E&DeviceType=Desktop&LangRequested=False

[71] https://dig.watch/processes/global-digital-compact

[72] https://www.un.org/ai-advisory-body

emphasises the importance of transparency, accountability, and the equitable distribution of AI benefits.

The **UN Human Rights Council** addresses the implications of AI on human rights. It has conducted several sessions focusing on the impact of AI on privacy, freedom of expression, and non-discrimination. The Council advocates for AI governance frameworks that protect human rights and ensure that AI technologies are used ethically and responsibly.

Group of Governmental Experts on Lethal Autonomous Weapons Systems (UN GGE on LAWS) is a specialised body that examines the implications of AI in military applications, particularly autonomous weapons systems. The group aims to develop international norms and regulations to prevent the misuse of AI in warfare and ensure compliance with international humanitarian law.

The **International Telecommunication Union (ITU)** is a specialised agency of the UN responsible for issues related to information and communication technologies. The ITU has been actively involved in AI governance through its AI for Good Global Summit, which brings together stakeholders from various sectors to discuss the ethical, technical, and societal implications of AI. The ITU also works on standardisation efforts to ensure interoperability and security in AI systems.

The **International Labour Organization (ILO)** addresses the impact of AI on the workforce, focusing on issues such as job displacement, skills development, and workers' rights. The ILO advocates for policies that ensure a fair transition for workers affected by AI and promotes the use of AI to enhance job quality and productivity. The organisation also collaborates with member states to develop strategies for managing the labour market implications of AI.

The **World Health Organization (WHO)** examines the use of AI in healthcare, focusing on areas such as diagnostics, treatment, and public health. AI technologies have the potential to revolutionise healthcare

delivery, but they also raise ethical and regulatory challenges. The WHO develops guidelines and best practices for the safe and effective use of AI in healthcare, ensuring that these technologies benefit all populations.

Organisation *for Economic Co-operation and Development (OECD)*

The OECD has developed the OECD AI Principles, which provide a framework for the responsible development and deployment of AI. These principles emphasise the importance of human-centred values, transparency, robustness, and accountability. The OECD also conducts research and provides policy recommendations to its member countries to help them navigate the complexities of AI governance (OECD AI Principles).

European *Union (EU)*

The European Union has been proactive in establishing comprehensive AI governance frameworks. The EU AI Act, proposed in 2021 and approved in June 2024, aims to create a regulatory environment that balances innovation with the protection of fundamental rights. The Act categorises AI systems based on their risk levels and imposes stringent requirements on high-risk AI applications. The EU also collaborates with international partners to harmonise AI regulations and promote global standards.

International Organization for Standardization (ISO)

The ISO develops international standards for AI to ensure interoperability, safety, and reliability. The ISO/IEC JTC 1/SC 42 committee is specifically dedicated to AI standardisation, covering areas such as AI terminology, trustworthiness, and governance. These standards provide a common framework for organisations worldwide to develop and deploy AI systems responsibly (ISO/IEC JTC 1/SC 42).

Financial Action Task Force (FATF)

The FATF examines the implications of AI for anti-money laundering (AML) and counter-terrorist financing (CTF) efforts. AI technologies can

enhance the detection and prevention of financial crimes, but they also pose new risks. The FATF develops guidelines and best practices for the use of AI in AML/CTF, ensuring that these technologies are used effectively and responsibly.

International Monetary Fund (IMF)

The IMF explores the economic implications of AI, focusing on areas such as productivity, employment, and inequality. The organisation conducts research and provides policy advice to member countries on how to harness the benefits of AI while mitigating its potential risks. The IMF also collaborates with other international organisations to promote a coordinated approach to AI governance.

Council of Europe

In 2024, the Council of Europe adopted the Framework Convention on AI and Human Rights, Democracy, and the Rule of Law. The Convention, which opens for signature on 5 September 2024, aims to provide a comprehensive framework for AI regulation, ensuring that AI technologies are developed and used in ways that respect fundamental rights and freedoms. Although the Council of Europe is a regional body, the negotiation of the Convention also involved several non-European states (Argentina, Australia, Canada, Costa Rica, the Holy See, Israel, Japan, Mexico, Peru, USA).

G7

In the framework of G7, the Hiroshima AI Process was launched in May 2023 to promote the establishment of international rulemaking for advanced AI systems. Within this process, three key documents were developed in 2023:

- A set of International Guiding Principles for All AI Actors, meant to be applied by all AI actors, "when and as relevant and appropriate, in appropriate forms, to cover the design, development, deployment, provision and use of advanced AI systems".

- A set of guiding principles and a code of conduct for organisations developing advanced AI systems. These relate to issues such as the evaluation and mitigation of risks across the AI lifecycle, the identification and mitigation of vulnerabilities and incidents, the development and disclosure of AI governance and risk management policies, and the development and deployment of content authentication and provenance mechanisms.

2.4.4 Bilateral Cooperation

There is an increasing number of regional initiatives. This cooperation is particularly well developed between the United States and the European Union through the Trade and Technology Council established by two sides. They discuss and coordinate their approaches to AI governance and standardisation.

2.4.5 Regional Cooperation

The work done at a European level (EU and Council of Europe) on AI governance has already been mentioned above. But there are also other regional organisations exploring AI policy and governance issues.

One example is ASEAN, which came up with a Guide on AI Governance and Ethics (2024) providing guidance for organisations across the region that design, develop, and deploy AI technology in commercial, non-military, or dual-use applications. The guide also includes recommendations on national and regional initiatives that ASEAN governments can consider implementing to design, develop, and deploy AI systems responsibly.

The African Union is also exploring options for AI governance mechanisms, as illustrated by a draft white paper by AUDA-NEPAD. Moreover, in 2021, the African Commission on Human and Peoples' Rights adopted resolution 473/2021 which calls on state parties to work towards comprehensive legal and ethical governance frameworks for AI

technologies to ensure compliance with the African Charter and other regional treaties. In mid-2024, the African Union adopted the Continental Artificial Intelligence Strategy titled 'Harnessing AI for Africa's development and prosperity'.

2.4.6 Public-private partnerships

Global Partnership on AI (GPAI)

The GPAI is an international initiative that brings together experts from governments, industry, academia, and civil society to promote the responsible development and use of AI. The partnership focuses on four key areas: responsible AI, data governance, the future of work, and innovation and commercialization. GPAI facilitates knowledge sharing and collaboration among its members to address global AI challenges.

World Economic Forum (WEF)

WEF, with its annual gathering in Davos, is one of the primary spaces for public-private partnerships. WEF has put AI high on the agenda. In addition to numerous AI-related sessions in Davos, WEF has the following ongoing mechanisms for dealing with AI:

The WEF's Platform on AI and Machine Learning serves as a cornerstone for its AI governance activities. This platform brings together actors from both the public and private sectors to co-design and test policy frameworks, with a focus on the following key areas:

- Standards for protecting children: Developing guidelines to ensure that AI technologies are safe for children.
- AI regulators for the twenty-first century: Creating a regulatory framework that is adaptable to the rapid advancements in AI.
- Facial recognition technology: Addressing the ethical and privacy challenges posed by facial recognition systems.

- The Global AI Council is the second main AI mechanism at WEF, established to provide policy guidance with a focus on the following policy areas:
- Ensuring that AI technologies are safe and secure.
- Promoting ethical AI practices that align with global values.
- Exploring the implications of machine learning and predictive systems on global risks and international security.

2.4.7 Creating new multilateral mechanisms for AI governance

In addition to utilising existing organisations, various proposals exist to create new AI governance mechanisms. In the search for the design of such new mechanisms, frequent analogies are made to the following organisations and initiatives: the International Panel on Climate Change (IPCC), and the International Atomic Energy Agency (IAEA).

The International Panel on Climate Change (IPCC)

The IPCC was established in 1988 by the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP) to provide a clear scientific view on the current state of knowledge in climate change and its potential environmental and socio-economic impacts (CFR, 2023). The IPCC's reports are instrumental in shaping international climate policies. They serve as a scientific basis for negotiations and agreements, such as the Paris Agreement.

There are the following analogies with IPCC that could inspire new AI mechanisms:

Scientific Basis and Peer Review: The IPCC's rigorous assessment process ensures that its reports are credible and widely accepted. Similarly, AI governance can benefit from a structured peer-review process to validate the ethical and technical aspects of AI systems (Will science diplomacy survive?).

Probabilistic Framing: The IPCC uses probabilistic terms (very likely, likely, unlikely) to describe the potential impacts of climate change, which helps in setting realistic expectations. AI governance can adopt a similar approach to frame the risks and benefits of AI technologies, thereby facilitating more informed decision-making (Will science diplomacy survive?).

Multistakeholder Involvement: The IPCC involves scientists, policymakers, and other stakeholders in its assessment process. This multistakeholder approach ensures that diverse perspectives are considered. AI governance can similarly benefit from involving various stakeholders, including tech companies, governments, and civil society, to create a more balanced and comprehensive regulatory framework (São Paulo Multistakeholder Guidelines).

Transparency and Accountability: Transparency is a cornerstone of the IPCC's operations. All its reports and methodologies are publicly available, ensuring accountability. AI governance frameworks should also prioritise transparency, particularly in areas like data usage, algorithmic decision-making, and ethical considerations (AI diplomacy).

Policy Influence: The IPCC's reports have a significant impact on international climate policies. Similarly, a well-structured AI governance framework can influence national and international policies, ensuring that AI technologies are developed and used responsibly (UN 2.0 | Our Common Agenda).

International Atomic Energy Agency (IAEA)

The IAEA was established in 1957 as an independent organisation reporting to the United Nations General Assembly and the Security Council. Established to promote the peaceful use of nuclear energy and prevent its misuse, the IAEA's structure and operational principles offer valuable insights for developing effective AI governance frameworks.

The IAEA was used as inspiration for those whose primary AI governance concern is safety and risks. Typically, they highlight the following analogies between IAEA and new AI mechanisms:

Safeguards and verification: One of the most critical aspects of AI governance is ensuring that AI technologies are used responsibly and ethically. Similar to the IAEA's safeguards and verification mechanisms, AI governance could involve regular audits and assessments to ensure compliance with ethical standards and regulations, including algorithmic audits, data privacy assessments, and transparency reports.

Safety and security: The IAEA's focus on safety and security can be mirrored in AI governance by developing comprehensive safety standards for AI systems. This could involve risk assessment frameworks, incident reporting mechanisms, development of safety standards for AI.

Transparency and accountability: The IAEA's emphasis on transparency and accountability can be applied to AI governance by: promoting open data initiatives to ensure that AI systems are transparent and their decision-making processes can be scrutinised; establishing mechanisms to hold organisations accountable for the ethical and responsible use of AI, including penalties for non-compliance; encouraging public engagement and dialogue on AI governance to ensure that policies and regulations reflect the concerns and aspirations of society.

CERN (Conseil européen pour la recherche nucléaire)

CERN or the European Organisation for Nuclear Research, is one of the most prominent examples of international cooperation in the scientific field. In discussion on AI governance, the following analogies with CERN are frequently quoted:

Data governance: Just as CERN manages and preserves vast amounts of scientific data, AI governance must address the challenges of data management, including data privacy, security, and ethical use. The principles of transparency and openness in data handling, as practised by

CERN, are equally applicable to AI governance (European Organization for Nuclear Research).

Collaborative frameworks: CERN's collaborative model, which brings together diverse stakeholders from academia, industry, and government, can inspire similar frameworks for AI governance. The complexity and interdisciplinary nature of AI necessitate a collaborative approach to governance, involving technologists, ethicists, policymakers, and the public. This multi-stakeholder model can ensure that AI development is aligned with societal values and ethical standards (Policy Meets Tech - Journey Diary).

Standard setting and transparency: CERN's role in setting technical standards with full openness and transparency offers a valuable lesson for AI governance. The establishment of clear, transparent standards for AI development and deployment can enhance trust and accountability. Flexibility and adaptability: The flexible and adaptive governance structure of CERN is particularly relevant for AI, a field characterised by rapid innovation and change. AI governance frameworks must be designed to evolve in response to new developments and emerging risks. This requires a balance between regulatory oversight and the freedom to innovate, ensuring that governance mechanisms do not stifle technological progress (Keeping AI in check).

2.5 What are AI governance issues?

This section provides a survey of some key AI governance issues that diplomatic services have to address on an international level.

2.5.1 Socio-economic issues

AI has significant potential to stimulate economic growth. In production processes, AI systems increase automation, and make processes

smarter, faster, and cheaper, and therefore bring savings and increased efficiency.

Beyond the economic potential, AI can also contribute to achieving sustainable development goals (SDGs). For instance, AI can be used to detect water service lines containing hazardous substances (SDG 6 – clean water and sanitation), to optimise the supply and consumption of energy (SDG 7 – affordable and clean energy), and to analyse climate change data and generate climate modelling, helping to predict and prepare for disasters (SDG 13 – climate action). Across the private sector, companies have been launching programmes dedicated to fostering the role of AI in achieving sustainable development. Examples include IBM's Data and AI for Social Impact, Google's AI for Social Good, and Microsoft's AI for Good projects.

2.5.2 AI and environment

The environmental impact of AI technologies is a critical concern in international negotiations, primarily due to the substantial energy consumption and resultant carbon footprint associated with AI operations.

AI systems, particularly those involving deep learning and large-scale data processing, require significant computational power. Data centres, which house the servers and infrastructure necessary for AI operations, are major energy consumers. The carbon footprint of AI is further exacerbated by the energy mix used to power these data centers. The environmental impact is significantly higher in regions where fossil fuels dominate the energy grid. For instance, a study by the University of Massachusetts Amherst found that training a single AI model can emit as much carbon as five cars over their lifetimes (Strubell et al., 2019).

This highlights the urgent need for international cooperation to promote the use of renewable energy sources in powering AI infrastructure. AI technologies also contribute to environmental degradation through the extensive use of natural resources and the generation of electronic waste

(e-waste). The production of AI hardware, including servers, GPUs, and other components, requires significant amounts of rare earth metals and other non-renewable resources.

The extraction and processing of these materials have substantial environmental impacts, including habitat destruction, water pollution, and greenhouse gas emissions. Water usage is another critical environmental aspect of AI technologies. Data centres require substantial amounts of water for cooling purposes to prevent overheating of servers. This can lead to significant water consumption, particularly in regions already facing water scarcity.

Despite the environmental challenges posed by AI, these technologies also offer significant opportunities for environmental monitoring and protection. AI-driven systems can enhance the accuracy and efficiency of environmental data collection, enabling better monitoring of air and water quality, deforestation, and wildlife populations. Furthermore, AI can improve predictive capabilities for climate change by analysing vast amounts of environmental data to identify patterns and trends. This can help policymakers develop more effective strategies for mitigating the impacts of climate change and adapting to its effects.

Besides its promises, AI also brings challenges related to divides and inequalities. There are notable gaps in AI development and adoption, which can lead to increased societal divides if not properly managed. This is illustrated in the AI Preparedness Index (Figure 1) developed by the International Monetary Fund (IMF); the index shows that wealthier economies tend to be better equipped for AI adoption than low-income countries, many of which lack the infrastructure and skilled workforces needed to harness AI’s benefits. To ensure AI's benefits are widely shared, it is important to address these inequalities and prevent AI systems from reinforcing existing disparities or excluding people.

The disruptions that AI systems could bring to the labour market are another source of concern. Many studies estimate that automated systems will make some jobs obsolete, and lead to unemployment. Such concerns have led to discussions about the introduction of a 'universal basic income' that would compensate individuals for disruptions brought on the labour market by robots and by other AI systems. There are, however, also opposing views, according to which AI advancements will generate new jobs, which will compensate for those lost without affecting the overall employment rates.

Nevertheless, ongoing policy discussions focus on the need to address the impact of AI technologies on labour rights and workforce dynamics. The International Labour Organization (ILO), for instance, has been proactive in this regard, with its Global Commission on the Future of Work recommending, among other issues, a 'human-in-command' approach to technology in support of decent work, an international governance system for digital labour platforms, and regulations to uphold algorithmic accountability in the world of work.

One point on which there is broad agreement is the need to better adapt education and training systems to the new requirements of the jobs market. This entails not only preparing new generations, but also allowing the current workforce to re-skill and up-skill itself.

2.5.3 Safety and security

AI applications in the physical world bring into focus issues related to human safety, and the need to design systems that can properly react to unforeseen situations with minimal unintended consequences. Self-driving cars are one often cited example, but other applications, such as autonomous drones, industrial robots, and AI-powered medical devices, also highlight the importance of robust and reliable AI systems. As these AI systems become more integrated into our daily lives, addressing safety

concerns becomes increasingly critical to prevent accidents, misuse, and other harmful outcomes.

AI also has implications in the cybersecurity field. In addition to the cybersecurity risks associated with AI systems (e.g., as AI is increasingly embedded in critical systems, they need to be secured against potential cyberattacks), the technology has a dual function: it can be used as a tool to both commit and prevent cybercrime and other forms of cyberattacks. As the possibility of using AI to assist in cyberattacks grows, so does the integration of this technology into cybersecurity strategies. The same characteristics that make AI a powerful tool to perpetrate attacks also help to defend against them, raising hopes for levelling the playing field between attackers and cybersecurity experts.

AI is also looked at from the perspective of national security. The US Intelligence Community, for example, has included AI among the areas that could generate national security concerns, especially due to its potential applications in warfare and cyber defence, and its implications for national economic competitiveness.

AI is also increasingly relevant in the context of (international) peace and security. Its application in military contexts, particularly through autonomous systems, raises significant ethical and legal concerns. Autonomous weapons systems capable of making life-and-death decisions without human intervention pose risks such as unintended escalations, violations of international law, and threats to civilian populations. On the other hand, AI also offers positive potential in conflict management and mediation. For instance, AI can aid peacebuilding efforts by analysing extensive datasets to identify early warning signs of conflict. It can optimise resource allocation for humanitarian aid and support negotiation processes through real-time data analysis and scenario modelling.

2.5.4 Human rights and ethical concerns

AI systems work with enormous amounts of data, raising significant concerns regarding privacy and data protection. Online services such as social media platforms, e-commerce stores, and multimedia content providers collect extensive information about users' online habits and use AI techniques like machine learning to analyse the data and "improve user experiences". AI-powered products such as smart speakers also process user data, some of it of a personal nature, and facial recognition technologies (FRT) embedded in public street cameras have direct privacy implications.

This extensive data collection prompts several critical questions: How is this data processed? Who has access to it, and under what conditions? Are users even aware of how extensively their data is used? To ensure that AI advancements do not come at the expense of user privacy, potential solutions include strong privacy and data protection regulations, enhanced transparency, and accountability for tech companies, and embedding privacy guarantees into AI applications during the design phase.

Beyond privacy, AI algorithms can impact other human rights. For instance, AI tools designed to automatically detect and remove hate speech from online platforms could negatively affect legitimate freedom of expression, as algorithms may incorrectly identify texts as hate speech. Complex algorithms and human-biased datasets can reinforce and amplify discrimination, particularly among disadvantaged groups. Concerns about ethics, fairness, justice, transparency, and accountability in AI decision-making are also prominent, as algorithms can replicate human biases. For example, FRT programs may exhibit racial and gender biases if they are trained predominantly on photos of males and white people. If law enforcement agencies rely on such biased technologies, it could result in discriminatory decisions, including false arrests.

Addressing AI ethics concerns involves combining ethical training for technologists with developing technical methods for designing fair, transparent, and accountable AI systems. Researchers are also working on developing AI algorithms that can explain their decision-making processes, which could enhance understanding and improvement of these algorithms, ensuring they operate more fairly and transparently.

2.5.5 Protection of Data and Knowledge

Data is where AI gets its main inputs from, and this is why it is sometimes called the 'oil' of the AI industry. Generative AI models, for instance, rely on vast training datasets, which can encompass personal information, articles, papers, books, and so much more. But the public knows little about what exactly goes into an AI model, and there are more and more calls for clarity on what—and whose—data and knowledge is used by the developers of these models.

Data-related challenges go beyond the privacy matters discussed above, and there are two other major sets of challenges. Firstly, copyright holders worldwide are questioning in courts the use of their creative work in the development of AI. Secondly, facing a limited data volume of internet content, and in an attempt to respond to privacy and intellectual property concerns, companies are using synthetic data generated by AI.

The question of protecting the intellectual property used by AI platforms started to gain momentum in 2023, as writers, musicians, the photography industry, media publications and others launched court cases against generative AI companies overusing their materials for training AI models.

The outcomes of these court cases may shape future regulations regarding intellectual property and AI. In the ongoing debate on whether and what to regulate, some argue against strict copyright protection, emphasising a need to allow data scraping to facilitate AI progress. Others are of the view that, analogous with copyright, a 'learn right' should be

established to govern how AI companies use content for training. At the international level, there are focused discussions on AI and intellectual property rights within WIPO.

Moving on to synthetic data, many AI models are based on internet content, considered to be the 'ground truth' for AI development. This initial input into AI models is already depleted. Faced with these limitations, AI companies have started using synthetic data: data generated by AI to reproduce the characteristics and structure of original data. Synthetic data is also seen as a way to deal with concerns related to privacy, intellectual property, bias, and other considerations that (may) have a legal dimension. But the use of such data also comes with challenges. For example, while synthetic data can be used to reduce bias, it can also lead to the exact opposite result – bias amplification.

As the volume of AI-generated content online continues to grow, companies that scrape the internet for data to train their AI models encounter another challenge: a decline in the quality and relevance of these models. Studies suggest that when new AI models are trained on content produced by other AI systems, they develop persistent flaws that cannot be corrected. This results in outputs that are increasingly inaccurate and lack diversity, becoming more uniform over time.

One starting point for dealing with such challenges could be for regulators to put more pressure on AI companies to be transparent about the data they use to develop their models.

2.6. Conclusion: The future of AI diplomacy

As artificial intelligence continues to evolve at a rapid pace, the need for effective global governance and diplomacy becomes increasingly critical. This analysis has explored the multifaceted landscape of AI

diplomacy, examining why it matters, who the key players are, how it is conducted, where it takes place, and what the main issues are.

Several key themes emerge from this exploration:

- Complexity and interconnectedness: AI governance is a complex field that touches on a wide range of issues, from economic development and environmental sustainability to human rights and national security. This complexity necessitates a multidisciplinary and collaborative approach to diplomacy.
- Multiple stakeholders: Effective AI diplomacy involves a diverse array of actors, including national governments, tech companies, international organisations, and civil society groups. Balancing the interests and perspectives of these various stakeholders is a key challenge for diplomats and policymakers.
- Evolving mechanisms: AI diplomacy is still in its early stages, and new mechanisms for international cooperation and governance are continually being developed and refined. From existing multilateral forums to proposed new institutions, the landscape of AI diplomacy is dynamic and evolving.
- Balancing innovation and regulation: A recurring theme in AI diplomacy is the need to balance technological innovation with appropriate safeguards and regulations. Finding this balance requires nuanced understanding and careful negotiation.
- Global implications, local impacts: While AI is a global phenomenon, its impacts are often felt at local and national levels. AI diplomacy must therefore navigate between global standards and local implementations.

Looking ahead, several key challenges and opportunities are likely to shape the future of AI diplomacy:

- Bridging the AI divide: As AI technologies continue to advance, there is a risk of widening global inequalities. Diplomatic efforts will need to

focus on ensuring equitable access to AI benefits and building AI capacity in developing nations.

- Addressing emerging risks: As AI systems become more powerful and autonomous, new safety and security risks may emerge. International cooperation will be crucial in identifying and mitigating these risks.
- Ethical AI frameworks: Developing globally accepted ethical frameworks for AI development and deployment will be a key focus of future diplomatic efforts.
- Data governance: Issues surrounding data privacy, ownership, and cross-border data flows will continue to be central to AI diplomacy.
- AI in conflict and peacebuilding: The dual-use nature of AI in military and peace applications will require careful diplomatic navigation.

In conclusion, AI diplomacy is set to become an increasingly important field in international relations. As AI technologies continue to advance and permeate various aspects of society, the need for effective global governance will only grow. Success in this arena will require innovative diplomatic approaches, cross-sector collaboration, and a commitment to ensuring that AI benefits humanity as a whole. The foundations laid today in AI diplomacy will play a crucial role in shaping the global AI landscape of tomorrow.

3

THE BATTLE OF INFLUENCE ON AI GOVERNANCE

Roxana Radu, United Kingdom / Romania

3.1 The current state of affairs

The development of artificial intelligence (AI)[73] dates back to the 1950s, but it has never been as popular as it is today. While the historical breakthroughs have been on "narrow intelligence" models, today we see multi-modal AI using a range of sources/inputs to improve general-purpose technology. For instance, based on neural networks and pattern recognition, deep learning became increasingly successful when natural language processing integrated the transformer architecture in 2017, allowing large language models to advance very quickly. Generative AI tools, such as OpenAI's ChatGPT launched in November 2022, are seeing mainstream adoption across the public and private sectors.

Today, AI is not artificial general intelligence, meaning that machines do not have their own form of "superintelligence". As a general-purpose

[73] *Roxana Radu* is an Associate Professor of Digital Technologies and Public Policy at the Blavatnik School of Government and a Hugh Price Fellow at Jesus College. This article originally appeared in the 2024 edition of the Geneva Policy Outlook.

technology, however, AI is used for a wide range of purposes, from industrial applications to large language models that generate new text, images or videos. The fast developments in machine learning in the last 5 years show the great potential of this powerful technology, which can also cause disastrous consequences. The future of AI will be built on models that move from facial to emotional recognition in seamless ways, incorporating more data systems than ever before. These advances mostly come from private research labs, generally funded by gatekeeper companies such as Microsoft, Google, or Meta in Silicon Valley, or Alibaba, Baidu or Tencent in China. These companies have invested significantly in both data collection and computing power in the last decade.

Given the growing geopolitical division, AI is developing in an era of uncertainty. 2024 will be a record year for the number of elections held around the world, bringing more than 2 billion people to the polls, including in the United States, India, Mexico and South Africa. With AI governance as a top policy priority, changes in these administrations will impact global developments in the field. Second, in the context of the technological competition between the US and China, Taiwan is an essential source for the global supply chain of advanced graphical processing units (GPUs), which are central to the development of AI models. There could be escalatory scenarios in which the US and China take measures and countermeasures against one another, while their allies ponder on the best possible alignment. Third, the search for checks and balances to counter the unprecedented power of a few players in the AI industry will define regional and national governance arrangements. Contestation will emerge from these three undercurrents that influence the development of AI.

3.2 The battle for influence over AI's global rulebook

The international community acknowledges the risks and complexities of governing AI. There are many risks amplified by the use of AI, ranging from relatively narrow threats to broader societal risks. At one end of the spectrum, AI systems produce biased, erroneous inferences, or even "hallucinations" presenting false or misleading information as a fact. On the other end, they trigger structural and societal changes via job losses and educational or environmental impacts. While recent discussions tend to focus more on existential risks and potential loss of human control, the everyday harms engendered by AI use need full consideration. A growing disconnection between AI and digital inclusion makes it visible that countries and populations are already suffering from both digital divide(s) and digital poverty.

We are witnessing various efforts to govern AI. First, on the multilateral level, several processes have started: the Council of Europe's negotiation of a legally binding instrument on the development, design and application of AI systems. Meanwhile, the UN Secretary-General has appointed a High-Level Advisory Board on AI tasked to map the governance landscape and propose policy and institutional options ahead of the Summit of the Future in September 2024. A first global AI Safety Summit convened by the UK in November 2023 recently built momentum for bridging the gap in AI safety testing.

Second, discussions on regulation are taking place at the national level. Alongside the EU's ongoing work in respect of its AI Act, China's internal attempts to provide regulatory mechanisms and the more recent US Executive Order on AI, more and more governments in Latin America, Africa and Asia are discussing AI bills. What is clear is that the road towards governing and regulating AI cannot be a quick fix. Many countries are still figuring out their position in the regulatory debates. Instead, it is a long-term process to govern a general-purpose technology that is

already here, while allowing for sufficient flexibility to accommodate constant advances.

Despite these initiatives, key questions dominating the global AI discussions remain unanswered: what future-proof solutions can be designed on a governance level? What are the dangers of a fragmented governance system? How could a more coherent regime be optimised for the public interest? Are we ready to develop governance ecosystems that are centred on the well-being of humanity and remain mission-driven, rather than reactive ones that try to fix solutions to current problems?

3.3 The role of international Geneva

Geneva can play a significant role in AI governance. With a critical mass of international organisations and NGOs, Geneva is a global hub for both technical standardisation and international law – in particular humanitarian and human rights protections. More deliberate efforts for these established fields to inform AI developments are needed moving forward. The humanitarian ecosystem is rapidly transformed by AI and the data of an unprecedented number of civilians is now being handed over to machines, with long-term, but unknown consequences.

Yet, most organisations seem to work in their silos, which makes it difficult to speak with a strong voice in global AI conversations. The majority of international institutions based in Geneva have already embarked on the AI journey, either by applying AI tools internally or by contributing to or building AI-enhanced instruments for global benefit. Directly and indirectly, Geneva is a part of how consequential AI decisions are taken.

Additionally, active multi-stakeholder networks based in and around Geneva remain at the forefront of the "AI for good" movement. Yet, the meaning of 'good' is not always clear and not commonly shared

by all stakeholders. Previous attempts to draft ethical principles in the AI world have proven their limitations, as they do not travel globally. In this light, two pathways are worth pursuing: first, agreeing on the central cornerstones that different stakeholders value and using the convening power of International Geneva to find a common vocabulary. Second, International Geneva can facilitate discussions about mechanisms for the protection of non-digital knowledge from AI-related disruptions.

Finally, academia remains an underrepresented stakeholder in the AI governance conversations. While some universities advance the field of AI, there is an imbalance in the academics' access to discussion venues, which can be minimised in International Geneva. A growing number of governments have announced public funding and pooling of resources for national supercomputers (a recent example is the Swiss National Supercomputing Centre), aimed at minimising the gap between private and public expertise. Future discussions should not only integrate the evidence and views provided by academic institutions but also actively inform future research on societal transformations engendered by AI. The time is now.

4

TACKLING THE IMPACT OF ARTIFICIAL INTELLIGENCE ON HUMAN RIGHTS AT THE UNITED NATIONS

HUMAN RIGHTS-BASED AND "COMPREHENSIVE" EFFORTS

Anna Walch, Austria

4.1 Introduction

Artificial intelligence [74] has become a dominant issue in human rights fora – and beyond – at the UN. As noted in a recent Office of the High Commissioner on Human Rights (OHCHR) report[75], an astonishing number of at least 135 reports from the Special Procedures of the UN Human Rights Council discuss aspects of digitalization. A quick scan of these reports suggests that at least 15 of them specifically examine the impact of artificial intelligence on human rights.

---

[74] *Anna Walch* is Austrian Diplomat and working at the Mission of Austria to the United Nations in Geneva. All views expressed in this article are the personal views of the author only.

[75] Office of the High Commissioner for Human Rights, A/HRC/56/45, "Mapping report: human rights and new and emerging digital technologies", 2024, p.2

They deal, among other, with questions such as the potential racial bias that artificial intelligence and algorithms may have; how to ensure that new technologies such as artificial intelligence and automation do not have disproportionate adverse impacts on women's human rights; artificial intelligence and the rights of persons with disabilities; the implications of artificial intelligence technologies on the rights to freedom of opinion and expression, privacy and non-discrimination; the complex challenges artificial intelligence, hacking tools and digital identification are posing to association and assembly rights; challenges to judicial independence; the impact of assistive and robotics technology, artificial intelligence and automation on the human rights of older persons; the specific risks stemming from the use of artificial intelligence by police and security forces for persons belonging to minorities; the impact of artificial intelligence on the human rights of those living in poverty. It is fair to say that most of the 46 thematic mandate holders of the Human Rights Council have commented on the impact of AI on their field of competence.

At the UN, AI is the new sexy. The debate over the United Nations' role in AI governance has already led to a proliferation of initiatives by various actors and a remarkable rise in discussions related to AI across different UN fora and bodies in 2023 and 2024. The fault lines run along geopolitical and economic interests, values, and authority of interpretation, with human rights at the core of many of the differing positions taken by governments in the UN's AI governance discussions.

The UN General Assembly and the Human Rights Council have both adopted resolutions, by consensus, which collectively delineate the member states' visions of the governance of artificial intelligence. The initial resolution pertaining to AI was adopted by the Human Rights Council in July 2023HRC Resolution. This resolution (which will be discussed in detail below) established a number of principles for the development and utilization of AI in a manner that is consistent with international human

rights. In November 2023, the General Assembly adopted a resolution presented by Czechia, Mexico, the Maldives, the Netherlands, and South Africa[76] on the "Promotion and protection of human rights in the context of digital technologies", which repeated and elaborated on several of the principles established by the Human Rights Council.

Early in 2024, the United States[77] submitted the first proposal for a General Assembly Resolution exclusively on the issue of artificial intelligence, entitled "*Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*"[78]. The issue of human rights proved to be a significant point of contention during the negotiations.

The text was adopted in March 2024 after lengthy discussions and celebrated as a landmark resolution.

However, this was not the conclusion of the matter. Just a few weeks after the US resolution was adopted, China announced its intention to *'balance out the overly human rights-heavy approach with another resolution exclusively on the development aspect of artificial intelligence*'. The text entitled "*Enhancing international cooperation on capacity-building of artificial intelligence*"[79] did initially not mention human rights at all. References to human rights were included in several

[76] Czech Foreign Ministry. Dec. 2023. The UN General Assembly adopts the first initiative on human rights and digital technologies, https://mzv.gov.cz/jnp/en/issues_and_press/press_releases/the_un_general_assembly_adopts_the_first.html

[77] https://www.state.gov/united-nations-general-assembly-adopts-by-consensus-u-s-led-resolution-on-seizing-the-opportunities-of-safe-secure-and-trustworthy-artificial-intelligence-systems-for-sustainable-development/

[78] UN Ass. A/78/L.49; Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development, https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf

[79] UN General Ass., A/78/L.86; Enhancing international cooperation on capacity-building of artificial intelligence, https://documents.un.org/doc/undoc/ltd/n24/183/80/pdf/n2418380.pdf

paragraphs after difficult negotiations which ultimately allowed for a consensual adoption[80].

The accelerated emergence of disparate resolution initiatives pertaining to a singular topic, a phenomenon notably uncommon in the UN context, exemplifies the intense competition for interpretative authority regarding the role of the United Nations in the governance of artificial intelligence. Each state that presented a resolution sought to shape a novel global understanding of AI before the prospective establishment of a global framework for digital cooperation – in the form of the 'Global Digital Compact'.[81]

4.2 Commencing discussions in the Human Rights Council

The Human Rights Council, a UN organ composed of 46 elected member states, convenes three times a year to address human rights issues and situations. During these meetings, the Council negotiates an ever-growing number of resolutions that are adopted either by consensus or through a vote at the end of each session. Typically, these resolutions are tabled by groups of member states whose (foreign) policy priorities they reflect.

In 2012, the Human Rights Council settled the question of the necessity for a distinct human rights regime in regard to the digital space when

[80] Lederer, Edith M. UN adopts Chinese resolution with US support on closing the gap in access to artificial intelligence; July 2024, https://apnews.com/article/un-china-us-artificial-intelligence-access-resolution-56c559be7011693390233a7bafb562d1

[81] The Global Digital Compact, UN Office of the Secretary-Gen.'s Envoy on Technology, https://www.un.org/techenvoy/global-digital-compact

it declared unanimously[82] that: "[The Human Rights Council] *affirms that the same rights that people have offline must also be protected online, in particular freedom of expression* […][83]. This was groundbreaking. It paved the way for conversations about human rights and AI to concentrate on the application of existing human rights to AI, rather than debating whether they are applicable in the first place or whether a new normative framework would be required.

The quote above stems from the abovementioned 2012 resolution entitled "*The Promotion, Protection and Enjoyment of Human Rights on the Internet*", which was initiated by Brazil, Nigeria, Sweden, Tunisia, Turkey and United States of America[84]. In 2013, Brazil and Germany[85] introduced another text titled "*The Right to Privacy in the Digital Age*", another resolution that has become increasingly relevant in the era of AI. However, it was not until 2018 that informal discussions began to revolve around the issue of artificial intelligence, and the potential need for the Human Rights Council to address more specifically the human rights implications of newer emerging technologies.

The conversation originated in the Advisory Committee of the Human Rights Council, which had been established to function as a "think-tank" for the Council and works at its direction. The Austrian Chair of the Committee addressed a letter to the president of the Human Rights Council,

---

[82] China and Cuba voiced oral reservations during the adoption. A reservation has no legal implications.

[83] A/HRC/RES/20/8, The promotion, protection and enjoyment of human rights on the Internet, 16 July 2012

[84] Turkey left the group in 2018

[85] Brown, D. Nov. 2013. Brazil, Germany introduce resolution on Right to Privacy in the Digital Age at UN General Assembly, https://www.accessnow.org/brazil-germany-introduce-resolution-on-the-right-to-privacy-in-the-digital/

suggesting mandating the Committee with a study on the human rights implications of emerging technologies.

The Republic of South Korea was the first to take the decision to table a resolution. Austria, which was holding the Presidency of Council the European Union at the time and had prioritized digitalization[86], quickly joined the core group. They were then joined by Denmark Singapore as well as Brazil and Morocco, which at the time was in the process to develop a strategy to become a leader in AI on the African continent[87].

The first resolution tabled by the new core group was technical, asking for a report on the "*Human Rights Implications of New and Emerging Digital Technologies*"[88]. The mandate was deliberately kept broad, although AI, already at this stage, clearly was at the center of member states' discussions.

The initial follow-up resolution to the resolution[89] remained rather general, but it did establish several key concepts (see below). The cross-regional core group behind the resolution represented the divergent views of UN Member States on the governance of emerging technologies. The careful balancing of interests within the group facilitated the solicitation of support from the other member states once the text was submitted to

---

[86] Programme of the Austrian Presidency. Presidency of the Council of the European Union, July-Dec. 2018; https://www.eu2018.at/dam/jcr:52862976-3848-403e-a38a-6aac8bcbe34d/Programme%20of%20the%20Austrian%20Presidency.PDF

[87] La genèse du GITEX AFRICA Morocco, ADD, Sept. 2021; https://www.add.gov.ma/gitex-africa-morocco

[88] A/HRC/47/52, Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights, Report of the Human Rights Council Advisory Committee, 19 May 2021

[89] A/HRC/RES/47/23, New and emerging digital technologies and human rights, 16 July 2021

the wider UN membership. The core group favored a consensus-based approach and a gradual strengthening of the language to prevent getting caught up in growing geopolitical tensions surrounding AI[90].

Finally, in July 2023, the Human Rights Council adopted the first UN resolution establishing several principles on human rights and artificial intelligence systems[91]. It recognizes that to harness the benefits of artificial intelligence for humanity, adequate human rights safeguards must be in place and that without such safeguards, artificial intelligence systems can entail serious risks for the protection, promotion, and enjoyment of human rights.

The text emphasizes that AI systems are prone to embedding and exacerbating bias, potentially leading to discrimination and inequality, to intensify threats from misinformation, disinformation and hate speech. Importantly, member states agreed that certain applications of artificial intelligence "*present an unacceptable risk to human rights*."[92]

The resolution then sets out a list of requirements under existing international human rights law standards to ensure that human rights and fundamental freedoms are respected, protected and promoted throughout the AI life cycle. These include, inter alia, introducing frameworks for human rights impact assessments, exercising due diligence to assess, prevent and mitigate adverse human rights impacts, and ensuring that effective remedies are available. It also requires human oversight, accountability, and legal responsibility.

The resolution establishes that data used in the training of algorithms must be accurate, relevant, representative and audited against encoded

[90] Nonetheless, resolution A/HRC/RES/47/23 was brought to a vote by China (in which China, Venezuela and Eritrea abstained, all other members of the Human Rights Council voted in favor); see: A_HRC_47_2.docx (live.com), p. 69

[91] A/HRC/RES/53/29, New and emerging digital technologies and human rights, 18 July 2023

[92] A/HRC/RES/53/29, p. 3, PP22

bias to protect individuals from discrimination. This, so the text continues, requires transparency and *explainability* of artificial intelligence-supported decisions, considering the various levels of human rights risks arising from these technologies.

It should be noted that the resolution is consistent with some of the principles found in the European Union Artificial Intelligence Act – which was still under negotiation when the resolution was adopted – and places them in the context of human rights obligations. This includes the risk-based approach, where riskier AI applications require a higher level of transparency and *explainability* to ensure their compliance with human rights and the understanding, that certain AI applications present an unacceptable risk to human rights and can therefore never be commercialized or used in a manner that is compliant with human rights[93].

It is also noteworthy that the resolution enshrines commitments regarding AI systems not only based on human rights but deduces them from *the inherent dignity of the human person* – i.e., directly from the source of all human rights. Artificial intelligence, especially when integrated into decision-making systems with some level of autonomy, challenges the essence of the human condition: the recognition that a person must be able to exercise control over the decisions that technology may make on their behalf,[94] and the right to mental integrity. The Universal Declaration of Human Rights, by deducing all human rights from the inherent dignity of the human person, provides a strong foundation for protecting human decision-making and human agency. By including a reference to the dignity of the human person, the resolution provides for the

[93] One can think of Lethal Autonomous Weapon Systems or AI-based social scoring systems.

[94] See also: Inverardi, The Challenge of Human Dignity in the Era of Autonomous Systems; p. 25-29, In: Werther/Prem/Lee, Perspectives on Digital Humanism, 2022

consideration of emerging issues related to recent developments in the field of artificial intelligence within the existing framework of the Universal Declaration of Human Rights and subsequent human rights treaties.

4.3. The first big question: Can we talk about the "r"-words (human rights, risks and regulations)?

At the core of the resolution are two of the most carefully calibrated paragraphs, one stating "*that artificial intelligence systems, when adequate human rights safeguards are in place, have potential for the promotion, protection and enjoyment of human rights* "[95] and the second one "*that artificial intelligence systems, when used without appropriate safeguards […], can entail serious risks for the protection, promotion and enjoyment of human rights*."[96]

The duality of new and emerging technologies – offering both opportunity and threats simultaneously – has likely existed since the origins of technological innovation itself. This duality is central to many of the discussions that UN Member States are having about AI. There are multiple layers to this discussion and the intersecting positions taken by states that can roughly be divided into *three groups*:

The *first group* of countries has a developed AI industry and is skeptical about regulation, driven by economic interest and/or ideology. These countries often argue that technology itself is "neither good nor bad" and they generally prefer to attribute any potential downside – or "risk" – of AI to its inadequate use or "misuse".

The *second group* of countries lags in AI development and tends to believe that establishing human rights safeguards would prevent them

[95] A/HRC/RES/53/29, p. 3, PP 20
[96] A/HRC/RES/53/29, p. 3, PP 21

from catching up. These countries prefer to focus on the opportunities that come with AI in terms of the realization of human rights. It is important to acknowledge that, while this group has traditionally included many developing countries, there has been a notable shift in the discourse towards a much stronger focus on the potential risks that AI may pose to the right to non-discrimination, economic, social, and cultural rights, and the right to development.

The *third group* of countries has a somewhat developed AI industry but lags the major players and advocates for some degree of regulation, driven by values and/or economic interests. This group of countries tends to promote the development of robust human rights standards for AI within UN and other treaty-based fora and, more broadly, the enforcement of the rule of law in the digital space.

It is not surprising that the acknowledgment of the risks associated with artificial intelligence systems sparked heated discussions among countries at the Human Rights Council. While many countries argued for stronger language without caveats (i.e., AI systems always carry risks, even with safeguards in place), others called for a more cautious statement that would not prejudge the nature of the technology.

4.4 The second big question: When is the right time to consider human rights?

The Advisory Committee of the Human Rights Council, in its report on New and Emerging Technologies[97], proposed the concept of "*datafication cycle*" to move away from specific technologies and instead arrive

[97] A/HRC/47/52, Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights, Report of the Human Rights Council Advisory Committee, 19 May 2021

at general statements about the implementation of human rights at each stage in the development of new technologies.

This was a significant point, given that at the outset of the debate, it was frequently asserted that new technologies were evolving at an accelerated pace, resulting in human rights regulatory initiatives consistently trailing behind the latest technological developments. This, it was argued, rendered such initiatives ineffective. The authors of the report, therefore, emphasized the necessity for guidance on human rights to be both specific and tailored to each stage of the datafication cycle. This, they argued, would enable the establishment of comprehensive standards for the development of new technologies.

Including the term "datafication cycle" in a UN resolution did not prove consensual. The dynamic in UN human rights fora is such that new – "technical" – concepts from other sectors are frequently subjected to criticism from countries that are opposed to the progressive development of human rights law.

The 2019 resolution on "New and Emerging Technologies" uses the formula "the conception, design, use, development, and further deployment of technologies" to describe the stages relevant to specific human rights guidance. A number of countries contested the inclusion of the term "conception", arguing that the creative process of invention and the act of imagining new applications may not be regulated or hindered by any restrictions. Nonetheless, "conception" remained in the text and was later included in the first resolution on the "Promotion and protection of human rights in the context of digital technologies "adopted by the UN General Assembly in November 2023. That resolution expands even further on the stages of tech-development by including "the conception, design, development, deployment, use, sale, procurement or operation"[98] of

[98] A/C.3/78/L.49/Rev.1, Promotion and protection of human rights in the context of digital technologies, 8 November 2023

technologies. The first resolution of the UN General Assembly on AI specifically in 2024[99] skips the conceptualization phase and delineates the "pre-design, design, development, evaluation, testing, deployment, use, sale, procurement, operation and decommissioning" as pertinent stages.

The 2023 Human Rights Council resolution on "New and Emerging Technologies" includes for the first time the concept of the "lifecycle" of artificial intelligence systems. This reflects an effort to examine the technology in a comprehensive and global manner. The General Assembly resolutions on digital technologies and on AI subsequently adopted the concept, as did the UN Secretariat in of the Global Digital Compact[100].

4.5 The third big question: Human rights-based or holistic?

Another key paragraph of the same resolution of November 2023 stresses:

> "[the] importance of a human rights-based approach to new and emerging digital technologies, taking into account States' obligations under international human rights law, a holistic understanding of technology and holistic governance and regulatory efforts".

Those states that sought to firmly anchor the regulation of digital technologies in human rights had long been advocating for a so-called "*human rights-based approach*". Other countries asserted during the negotiations that the regulation of new technologies cannot be based solely on human rights considerations. Instead, factors such as competitiveness and

---

[99] A/78/L.49, Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development, 11 March 2024

[100] See: https://www.un.org/techenvoy/global-digital-compact

The expression also appears in the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law as well as in the EU's AI act.

security must also be taken into account. A human rights-based approach, so these countries, would result in the undue prioritization of human rights over these other considerations.

These differing views led to a compromise that emphasizes "holistic" governance as well as "holistic" regulatory efforts. This leaves some room for interpretation on both sides of the argument regarding the hierarchy of governance objectives and where human rights come in.

The expression "*human rights-based approach*" has its origins in the UN development system. The Vienna Declaration and Programme of Action of 1993 asserts that democracy, development, and respect for human rights and fundamental freedoms are interdependent and mutually reinforcing. Following its consensual adoption, many UN agencies have embraced and expanded upon the concept of "a human rights-based approach" in their development cooperation. The human rights-based approach aimed to transition development efforts from acts of charity to the obligation of "duty-bearers" to respect, protect and fulfil the rights of "rights holders", who are empowered to claim their rights[101].

The Human Rights Council consensually adopted numerous resolutions on a human rights-based approach to various issues, such as combating human trafficking or reducing maternal mortality. Human Rights Council Resolution 32/13 on "The promotion, protection and enjoyment of human rights on the Internet" emphasized "*the importance of applying a comprehensive human rights-based approach when providing and expanding access to the Internet*". Although some countries had previously expressed discomfort with the concept, it had never been presented as a reason to call a resolution for a vote.

This changed in 2021, coinciding with the negotiations pertaining to the second iteration of the resolution on new and emerging technologies.

---

[101] UNSDG, "The Human Rights Based Approach to Development Cooperation, Towards a Common Understanding Among UN Agencies", 2003

The dynamics in the Human Rights Council had already been shifting for a couple of years. Several countries increasingly pushed back against progressive concepts and language on human rights, while advocating for a stronger focus on development.

The resolution was ultimately called for a vote[102]. During the public adoption of the resolutions, the delegation calling for a vote explained that it could not accept the term "human rights-based approach"[103]. The next iteration of the resolution could achieve consensus, including on the term "human rights-based approach".

The Human Rights Council's 2023 resolution is notable for being the only one within the UN on the issue of emerging technologies and AI to incorporate the concept of a human rights-based approach. In the UN General Assembly, attempts to establish this concept in resolutions were met with fierce opposition, resulting in its absence from the resolutions on AI or Digital Technologies, as well as from the Global Digital Compact.

4.6 What else can be done?

The "human rights-based approach"[104] continues to represent the gold standard for countries and other actors seeking to ensure robust human rights protection in the digital space, including in the context of AI. Those who espouse this approach consider it tantamount to the full realization of human rights across the entire spectrum of technological development.

---

[102] See: A/HRC/47/2, 9 December 2021, p. 69

[103] During the same session of the Human Rights Council a vote was also called on the resolution regarding "The promotion, protection and enjoyment of human rights on the Internet", which for the first time in ten years was not adopted by consensus. The vote call was also explained by the concept of the "human rights-based approach", which had been included in the text for many years.

[104] Other names are "human rights-centered" or "human rights by design".

This, they contend, will ensure the upholding of human rights in the context of AI. However, those who are critical of the concept argue that there is a lack of clarity surrounding the definition, which could result in excessive interpretation and unjustified restrictions for technology companies and research entities.

It is accurate to conclude that a definition is absent. Despite the concept of a human rights-based approach being extensively debated among stakeholders from civil society and academia[105], routinely employed by UN organizations[106] and incorporated into numerous strategy documents, no universally accepted definition has been endorsed by UN member states.

Current intergovernmental debates at the United Nations are characterized by the pursuit of an equilibrium within the perceived dichotomy of a development-oriented and a human rights-oriented approach. This dynamic is particularly salient in the realm of new technologies and AI. The 1993 consensus, which already postulated a symbiotic relationship – and interdependence – between development and human rights, is now facing challenge from developing countries, notably in the context of discussions on digital cooperation[107]. Some countries consider that the concept of a 'human rights-based approach' has been misused in the past by the UN development system and bilateral actors to condition development cooperation on the promotion of human rights norms[108].

---

[105] For example: What would a human rights-based approach to AI governance look like? – Global Partners Digital (gp-digital.org)

[106] For example: Digital Border Governance: a Human Rights Based Approach | OHCHR

[107] For example, the Group of 77 expressed their opposition to references to human rights in the open consultations on the Global Digital Compact, arguing that it should be a ‘development document’.

[108] LGTBI rights are the most controversial theme in this regard.

The evolution of human rights standards for the governance of AI must be considered within the context of these increasingly complex and challenging discussions. It would be optimal for a new global consensus to be established regarding the relationship between human rights and development aspects in the context of digital cooperation. This would entail clarifying that these are not mutually exclusive but rather mutually reinforcing. However, this will undoubtedly necessitate many more arduous discussions. The Human Rights Council – and the UN General Assembly – could advance the discussion and produce guidance for human rights-compliant AI regulation, inter alia by:

(1) Reaffirming the full applicability of international human rights law to AI and to other, specific emerging technologies

It bears repeating that international law and existing human rights norms and standards apply in the context of emerging technologies and digital activities. It would be beneficial to focus future discussions on specific applications of artificial intelligence that have a significant impact on people's lives. This could include intelligent devices, brain-computer interfaces, and advanced machine learning.

Furthermore, the field of quantum computing – and its applications in AI – offers a unique opportunity to establish consensus around human rights principles before the technology reaches market maturity. This approach could still prove beneficial in terms of facilitating the implementation of the technology in a way that supports sustainable development and the creation of value for the general public, as opposed to serving the interests of a select few major technology companies.

(2) Elaborate and agree on policy measures and practices that are required or recommended to uphold existing international human rights law throughout the life cycle of technologies, i.e., define a "human rights-based approach"

Human Rights Council resolution A/HRC/RES/47/23 highlights the importance of including a human rights perspective within standard-setting processes and bodies, as well as the necessity for these bodies to build their human rights expertise. Furthermore, the resolution promotes the transparency, openness and inclusivity of such processes and bodies.

By adopting a position on this matter, the Human Rights Council has initiated a discussion on human rights due diligence during a specific stage of the technological development process. It seems reasonable to suggest that the document has already had a significant impact, prompting experts in this particular policy area to consider the implications for their own work[109].

In the future, the most useful areas for further elaboration by the Human Rights Council would probably be to clarify the implications of a human rights-based approach at the different stages of the life cycle of new technologies, in the form of principles to be agreed upon by Member States. This may prove to be a challenging endeavor, requiring extensive deliberation and consensus building. Nevertheless, it is a demand that has been articulated by the technology sector.

(3) Mandate either the Office of the High Commissioner on Human Rights (OHCHR) or a new "mechanism" with further mainstreaming of human rights in AI related fora

In the future, the most significant impact will be to reinforce the OHCHR's capacity to function as a central repository of information on

[109] The adoption of the resolution initiated an exchange between the Office of the High Commissioner on Human Rights and the International Telecommunications Union (ITU). Member states at the ITU have picked up on the resolution and started a discussion about human rights in the intergovernmental bodies of ITU.

human rights in the context of technology. Scaling up the OHCHR's support to technology companies and to the mainstreaming of a human rights-based approach in technology-related fora would have a real-world impact.

Furthermore, there might be value in establishing a new "mechanism" of the Human Rights Council that would be tasked with advocating for the integration of human rights in the technology sector. Given the intricate nature of the existing institutional framework for digital governance, there seems to be a need for a multi-stakeholder platform that could facilitate the development of a taxonomy to bridge the gap between the human rights and the technical expert communities.

5

ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS: CATALYST OR CONUNDRUM?

Warren Bowles, South Africa

5.1 Introduction

The impact of Artificial Intelligence (AI)[110] on the lives of human beings, societies and the world can be described as ubiquitous. It pervades every aspect of human existence, and it is changing the face of work and industries globally. In doing so, AI is morphing at great speed. The innovations in which it is finding application are surpassing the boundaries of what used to be regarded as science fiction. As AI morphs and enhances innovation, it learns and develops to meet the needs of society. AI's neural networks, deep learning algorithms and language processing mimic what we would perceive to be "human". Its own metamorphosis is making it more user-friendly, efficient, and knowledgeable in aiding human

[110] *Warren Bowles* is a legal academic with over ten years of experience in both public and private sectors. An admitted Attorney of the High Court of South Africa, his research focuses on AI, Law, and Education. He also develops academic materials and assessments for law modules and contributes to academic governance and policy.

beings and societies with a variety of industry-specific tasks. In this way, AI is becoming an emphatic catalyst for change. It is paramount that it is used its use for the good of all humanity.

Although AI is a global phenomenon with unbridled potential to better the world, there have been instances of fallibility, bias, and inaccuracy. As part of its own learning and application, AI has been found to make mistakes based on *"made up"* information. AI-based Natural Language Processing (NLPs) can render speculative responses based on *hallucinations* or fake data. Moreover, there are reports of bias against certain groups based on defining human characteristics, whether these be race, gender or sexual orientation. This weakness in the armour of AI's formidable presence is the subject of much debate and in the quest to regulate it, rendering AI tools more trustworthy, safe, and accurate. These fallibilities present us with a conundrum; they evidence the great responsibility to ensure that AI is ethically used for the benefit of all.

This is the context in which this chapter grounds itself. To establish whether AI is a catalyst or a conundrum, we can resort to a Human-Rights framework. One of the many AI-related, ongoing debates centres Human Rights and their instruments. Human Rights, sometimes deemed "sacrosanct", are being contrasted with AI's technological prowess. This chapter aims to explore the catalyst versus conundrum dichotomy by looking at what AI is and the Human Rights framework in which it operates. It will also delve into the notion of risk versus innovation that lies at the intersection of AI and Human Rights. The article identifies a concrete set of selection of specific Human Rights that are affected by AI uses, and suggests recommendations for regulating AI within a Human Rights framework.

5.2 Towards understanding AI: Contextualising AI as the "human" in a digital world

AI as a technological phenomenon is defined in different ways. There is not a single definition that can properly capture what it is and what it can perform. John McCarthy coined the term "Artificial Intelligence" in 1956, as the general ability of machines to make decisions and choices mimicking human reasoning by going beyond their original programming. In some instances, it has been said that AI is able to make decisions on an analytical scale and at speeds that far exceed human capabilities. AI can be understood as is a machine's uncanny ability to think and learn like a human being. Whether AI processes language, or it processes data in order to make an assumption, observation or decision, it learns to do so based on what it learns from contexts that are shown to it through human input and intervention.

There are other expanded definitions of AI. For instance, the definition of the European Commission High-Level Expert Group on Artificial Intelligence (hereafter "AI-HLEG") defines AI by stipulating that it *"refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals*." This further emphasises that AI learns from context and stimuli, with limited autonomy. It is not completely autonomous in its learning because, at this present stage, it still requires human intervention to achieve a certain outcome or to achieve a certain goal, be it on the level of algorithm, design, data curation or AI use. However, they can be 'supervised' or 'unsupervised'. The AI–HLEG proposed defining goes further to say: *"AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."*

This elaboration of the AI – HLEG definition illustrates two elements of AI that make it display "human-like behaviour" digital world. In software form, AI can perform certain *human* functions, such as speaking, reading, researching, recognising and recalling objects (for example, faces) and constructing language. In hardware form, AI takes on certain *human* tasks, with autonomous vehicles and integrated in advanced robots. AI can imitate certain human traits and tasks, behaving as its own entity. However, AI-enabled tools still require human input in terms of curating datasets and surveying their performance. Although apparently "human" to a degree, there are concerns about AI possessing emotional intelligence and about its ability to demonstrate empathy for human beings.

There are several facets of what it means to be human. Namely, the ability to show multiple emotions, and to show empathy. According to Misselhorn, there are three criteria that define what genuine empathy entails:

- Congruence of feelings: empathy requires the person who empathizes to feel what it is like to *experience* other's emotions in a specific situation. This distinguishes empathy from a mere rational understanding of emotions.
- Asymmetry: the empathetic person feels a certain emotion solely because she mirrors another's sentiment in a situation distinct from her own. For this reason, empathy is not just a shared emotion like the shared joy of parents over the progress of their offspring, where the asymmetry-condition is not met.
- Other-awareness: There must be at least a rudimentary awareness that empathy is about the feelings of another individual. This accounts for the difference between empathy and emotional contagion, which occurs if one catches a feeling or an emotion like a cold. This happens, for instance, when kids start to cry when they see another kid crying.

Emotions and empathy are complex human characteristics. The concern with AI is that, by looking at the criteria above, it is not able to demonstrate empathy. An AI-enabled tool cannot experience *feeling* in the same phenomenological sense, thus not meeting the criteria above. What AI can do is respond to emotions by recognising facial expressions, vocal cues, physiological patterns, or affective meanings based on the way in which it has been trained. For instance, an AI-enabled chatbot can simulate empathic behaviour. It has been argued that AI, as part of its own machine-learning processes, demonstrates a certain degree of rationality to perform a task or arrive at a decision. At the outset, rationality concerns the understanding of patterns and regularities in one's life and it is an integral part of decision-making. It has been argued that there are *four types of rationality that determine how we make decisions. These are:*

1) Practical rationality: the focus of this rationality is on pragmatic and egoistic interests.
2) Theoretical or "intellectual "rationality: this rationality involves a conscious mastery of reality through abstract concepts.
3) Substantive rationality: this rationality considers a holistic view of past, present, and potential "value postulates" to make inferences.
4) Formal rationality: this rationality is based on predefined rulesets and formal procedures to find the "correct" solutions to specified problems.

In general terms, AI delivers its output and comes to its own decisions through algorithmic thinking and logic. This is one method of thinking rationally about a phenomenon or situation. Human rationality is complex and multi-dimensional. At times, it is also grounded on a foundation of moral principles. When looking at the type of rationality used by AI, the algorithmic orientation of these tools speaks more to formal rationality than to other types. The algorithmic methods of AI represent *"a finite*

sequence of well-defined, computer-implementable instructions… to solve a class of problems or perform a computation." This illustrates the limitations of AI when acting as a "*human*" in a digital world. Although AI demonstrates rationality in performing a task or deciding, it can only do so based human-generated data. It is the human input and curation of training datasets that ultimately informs any potential moral frame of reference, without AI adding any emotions of its own. Although Large Language Models (or LLMs) like Chat GPT can provide outputs that are logical and sensible, they do not develop a moral compass beyond the developers' parameters and/or what can be inferred from the training datasets. This raises several concerns about the limitations of AI interacting as a "*human*" in a digital world, when confronted with actual human users and their complex emotions, thoughts, attitudes, moral obligations, and social backgrounds. This is especially contentious when looking at AI in the context of Human Rights. There is a dialectical relationship between the quest their quest for bigger, better, and faster technological innovations, and the risks that they might present.

5.3 Innovation versus risk: Looking left and right at the intersection of AI and human rights

As a general premise, the word "innovation" is derived from the late Latin term *innovationem* (*innovatio* in its nominative form) which, if broken up into different parts, summarily translates into "a novel or new change". By extension, "introducing a new thing into an established arrangement."[111] This root definition of the term "innovation" demonstrates how something new, different and better has the potential to change the course of society that has gotten used to a particular *status*

[111] Online Etymology Dictionary (2024). Available at: https://www.etymonline.com/word/innovation (accessed 16 June 2024).

quo over a number of years. However, something new that has entered and changed the present establishment appears to be progress, but this has not come without its own pitfalls, shortcomings and risks. This is no different within the digitised and technological revolution in the world today, especially in the case of AI. Although groundbreaking and efficient, the risks presented by AI have become the subject of much controversy and debate especially within the context of human rights. At the left of this intersection are the *innovations* in which AI is being developed to benefit all humankind. To the right, we see the risks that AI presents for the protection and enforcement of Human Rights. Looking in both directions involves looking at how AI's abilities and potential can be properly balanced against the 'sanctity' (universality) and importance of human rights. This balancing exercise today is an innovation in itself and requires emphasis especially when it comes to the risks presented by AI in a variety of areas that impact on human life every day.

On a macro-level, concerns about where AI and human rights intersect have been looked at in literature from the point of departure that AI in the hands of authoritarian regimes present a variety of issues that can impinge on a variety of human rights. In any democracy or free society that is engaged with innovation and technological advancement, there are the potential risks that the misuse of AI may lead to harms such as decreased privacy, lack of accountability and entrenched biases. With regard to the impact on livelihoods, the negative impact of AI can unveil itself in the form of job replacement (causing extensive job losses), the automation of hiring decisions that are discriminatory in nature and even the notion of robots that would be designed to kill.[112] AI has also been seen to be used in the political arena, for example to attract voters, inform them and guide voters through the voting process. This is an example of how AI is permeating democratic society and the massive amounts of

[112] Ibid., 115.

data processed to predict trends, algorithms used to analyse peoples' activity and the messages conveyed in campaign messages present a dilemma for how AI can manipulate and change the course of democracy as well as the way that people engage in democracy.[113] Moreover, it is not only national governments and intergovernmental organisations that have a vested interest in the use of AI to achieve certain objectives – technology companies (for example, Google in the United States of America and Baidu in China), technical professional associations and civil society organisations have also become significant role-players in informing national and foreign policies of states in the AI frontier. It becomes more complicated in that, as states begin to explore their interests in AI, the field of AI is governed by a complex disconnected set of policies, directives, guidelines and strategies that have yet to be synced in accordance with how national governments and other role-players need to work together in using AI for their own benefit.[114] If one looks to Africa for example, the argument has been made that there is need for a legitimate and comprehensive framework that will guard against human rights abuses perpetrated as a result of AI in light of its potential to deregulate, denationalise and to threaten compliance with human rights standards.[115]

At a micro-level, when looking at AI in the context of individuals, AI has been able to make a positive impact on our daily lives in a variety of areas through data-driven decision-making that is quick, efficient and simple in areas such as hiring, determining credit scores and healthcare.

---

[113] Kouroupis, Konstantinos (2023). "AI and Politics: Ensuring or Threatening Democracy". In: Juridical Tribune 13:4. 582 – 583.

[114] Su, Anna (2022). "The Promise and Perils of International Human Rights Law for AI Governance". In: Law, Technology and Humans 4(2). 169.

[115] Ncube, Caroline B. Oriakhogba, Desmond O. Schonwetter, Tobias and Rutenberg, Isaac. *Artificial Intelligence and the Law in Africa* (2023). LexisNexis: South Africa. 68.

However, AI in this context presents with a variety of socio-ethical challenges in the forms of discrimination, unjustified action, privacy infringement, dissemination of disinformation, fallout from automation in job markets and safety concerns. These areas, among others, reveal AI's potential to infringe on various human rights as well as the moral and social values that they stand for.[116] By looking at discrimination as an example, there are concerns with AI that, although it functions more effectively than human beings by using algorithms to complete a task or automate a decision, it has been argued that these biases are perpetuated through AI by human beings themselves. The algorithms used by AI will therefore work with the biased input to render a decision unless there are safeguards in place to prevent such bias that informs the discrimination in question.[117] This can for example occur where discrimination with AI takes the forms of unequal access of certain (often marginalised) groups to these technologies and biased data which, for example, may not be a representative sample of a greater and more diverse dataset.[118] In the African context for example, this may be because the algorithms possibly emanate from an unrepresentative or incomplete training data set or the reliance on flawed information that contains historical inequalities, which can lead to a perpetuation of technological racism, inequality, injustice

[116] Aizenberg, Evgeni and van den Hoven, Jeroen (2020). "Designing for human rights in AI". In: Big Data & Society. DOI: Designing for human rights in AI - Evgeni Aizenberg, Jeroen van den Hoven, 2020 (sagepub.com) (accessed 30 May 2024).

[117] Chatterjee, Sheshadri, Sreenivasulu, N.S. and Hussain, Z. (2022) "Evolution of artificial intelligence and its impact on human rights from sociolegal perspective". In: International Journal of Law and Management. Vol 64(2). DOI: Evolution of artificial intelligence and its impact on human rights: from sociolegal perspective | Emerald Insight (accessed 30 May 2024).

[118] Nagy, Noémi (2023). "Humanity's new frontier': Human rights implications of artificial intelligence and new technologies". In: Hungarian Journal of Legal Studies 64(2). 240.

and marginalisation of Africans.[119] At both micro- and macro-levels, this serves as a good example of the misuse of AI and the negative impact that it can have on people and their human rights, especially in relation to rights based on human dignity, equality and freedom.

What is illustrated further in the instance of discrimination and bias is that, although AI may play only an instrumental role to perpetuate certain biases at the hands of human beings that give it such input, other concerns have been raised about AI being able to independently internalise these biases in its own deep-learning and algorithmic processes. This has been observed in literature to mean AI that thinks and operates entirely on its own, with a devious purpose such as being able to kill people or to infiltrate societies and take them over. However, it is apparent at this stage that the final product of artificial general intelligence (or AGI) only exists as a theoretical concept as opposed to a real phenomenon. AGI is then categorised as a type of "strong AI" which has the aim of creating intelligent machines that are entirely autonomous and independent of any human mind, but they would still have to, in a limited manner, learn and upgrade their abilities over time from various contexts. "Weak AI" on the other hand, is AI that is able to perform simple tasks based on a certain mindset of AI development and this is the type of AI that is the subject of various studies, debates and controversies today.[120] Therefore, we may not see the actual humanoid machine of tomorrow that has the capacity to safeguard or to infringe on human rights (depending on how it is used or influenced) but by looking at the AI that is currently innovating itself with human insight on the left of our intersection, making sure

[119] Ncube, Caroline B. Oriakhogba, Desmond O. Schonwetter, Tobias and Rutenberg, Isaac. *Artificial Intelligence and the Law in Africa* (2023). LexisNexis: South Africa. 60.

[120] IBM. "What is Strong AI?" (2024). Available at: https://www.ibm.com/topics/strong-ai (accessed 17 June 2024).

that the risks presented by it on the right of our intersection do not infringe on human rights must remain a priority to guard against the disregard and devaluation of human rights that may be committed by the machines possessing AGI in the future.

5.4 Human rights as a 'lens' in relation to AI

Human rights are premised, at the outset, on being guaranteed to human beings simply because they are "human". The notion of human beings possessing rights is multifaceted and is explained in different ways through different legal instruments. The preamble to the Universal Declaration of Human Rights (hereafter "UDHR"), opens with the following which provides insight into what human rights entail for the "human" element in its essence:

"Whereas recognition of the *inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world*"[121] (emphasis added).

Both the International Covenant on Civil and Political Rights (hereafter "ICCPR") and the International Covenant on Economic, Social and Cultural Rights (hereafter "ICESCR") open with the statement above in their preambles and it goes further to say that rights (those being human rights) "*derive from the inherent dignity of the human person*…"[122]

---

[121] Universal Declaration of Human Rights. Available at: https://www.ohchr.org/en/human-rights/universal-declaration/translations/english (accessed 19 June 2024).

[122] International Covenant on Civil and Political Rights. Available at: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights (accessed 19 June 2024) and International Covenant on Economic, Social and Cultural Rights. Available at: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights (accessed 19 June 2024).

(emphasis added) which speaks to the "human" and what is human in the term "human rights".

The UDHR goes further in Article 1 to state that:

> "All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood."[123] (emphasis added)

In a regional context, the European Union Charter of Fundamental Rights (hereafter "EUCFR") provides a value-based approach to what the conception of human rights means, the greater role that it is supposed to play in strengthening relations between states within the European Union and how human rights require protection in times of change, including with technology:

> "The peoples of Europe, in creating an ever-closer union among them, are resolved to share a peaceful future *based on common values*.
>
> Conscious of its spiritual and moral heritage, the Union is founded on the *indivisible, universal values of human dignity, freedom, equality and solidarity*; it is based on the principles of democracy and the rule of law. It places the individual at the heart of its activities, by establishing the citizenship of the Union and by creating an area of freedom, security and justice. …
>
> To this end, it is necessary to strengthen the protection of fundamental rights in the light of changes in society, social progress *and scientific and technological developments by making those rights more visible in the Charter*."[124] (emphasis added).

Human rights therefore stem from a variety of sources aligned to the ideals of the human condition and human existence – human dignity,

---

[123] Universal Declaration of Human Rights, op. cit.

[124] European Union Charter of Fundamental Rights. Available at: https://fra.europa.eu/en/eu-charter/title/preamble (accessed 19 June 2024).

equality, togetherness as people, universality in their application and freedom – and they are also based in this regard on values and principles. Their protection is also emphatically more important to ensure comity and peace among nations and their protection is more amplified in times of change and progress.

In viewing human rights as this "lens" through which a relationship between human rights and AI is to be formed, there are two sides of the same coin that become evident. The first side of the coin is that AI, in order to comply with various human rights and to even ensure their protection at the forefront of its development, must comply with certain principles in order to be trustworthy and responsible. For example, this would include the principles for the responsible stewardship of trustworthy AI as outlined in the *OECD Recommendation of the Council on Artificial Intelligence,*[125] such principles being:

- Inclusive growth, sustainable development, and well-being.[126]
- Respect for the rule of law, human rights and democratic values, including fairness and privacy.[127]
- Transparency and explainability.[128]
- Robustness, security, and safety.[129]
- Accountability.[130]

[125] OECD Recommendation of the Council on Artificial Intelligence. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 19 June 2024).
[126] S 1 (1.1).
[127] S 1 (1.2).
[128] S 1 (1.3).
[129] S 1 (1.4).
[130] S 1 (1.5).

In respect of the second principle above, the principle stipulates that AI actors (whether they are developers, deployers or users)[131] must respect the rule of law, human rights, democratic and human-centred values throughout the entire lifecycle of an AI system. This includes ensuring non-discrimination on various grounds in AI, remedying any instances of misinformation or disinformation and implementing the appropriate safeguards for AI which specifically includes human intervention, human supervision and redress for any misuse of the AI.[132] It is also worth noting that AI must not only operate in such a way that it respects and amplifies human rights, but any values, whether they are constitutional, moral or jurisprudential, that inform human rights must also be accounted for within the purview of what AI aims to achieve. This is the ideal to which a relationship between AI and human rights needs to strive, but some authors are concerned about how these principles are to be interpreted. This then leads into the second side of the coin. Several definitions of "fairness", for example, prevail and then the question is also asked about how "universal" human rights are around the world where cultural backgrounds, attitudes and traditions differ.[133] This means that any ambiguity in understanding the true aims of principles and how they are applied in real world scenarios may adversely affect the relationship between human rights and AI, in terms of the extent of any infringements on human rights that may occur.

In this regard, the *European Union Artificial Intelligence Act* (hereafter the "EU AI Act") notes that AI poses a threat to human rights and it

---

[131] "AI actors" are defined in the OECD Recommendation of the Council on Artificial Intelligence as: "those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI."

[132] S1 (1.2) (a) and (b).

[133] Su, Anna (2022). "The Promise and Perils of International Human Rights Law for AI Governance". In: *Law, Technology and Humans* 4(2), 170.

aims to provide high-level protection of fundamental rights by addressing the various sources of risk associated with AI using a clearly defined risk-based approach.[134] The EU AI Act does this by classifying AI into different categories, for example, prohibited artificial intelligence practices (usually AI that is entirely banned where it uses subliminal means to distort the behaviour of persons or enforces biases of certain vulnerable groups)[135] and AI systems that are high-risk (usually involving components of AI or where AI itself is a component of a third-party system).[136]With reference particularly to the EUCFR, the EU AI Act acknowledges that the development of trustworthy AI and the placing of appropriate obligations on various AI stakeholders will amplify and promote the protection of human rights such as the right to human dignity,[137] respect for private life and protection of personal data,[138] non-discrimination,[139] and equality between women and men.[140] This also seeks to, among other purposes, avoid hindering rights to freedom of expression,[141] and freedom of assembly.[142] This illustrates only a few of the human rights that can be adversely impacted by AI and what the EU AI Act aims to protect. For purposes of this contribution, this will be explored further in respect of two human rights, namely human dignity and freedom of expression.

---

[134] European Union Artificial Intelligence Act, item 3.5. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed 19 June 2024).

[135] Art 5 of the EU AI Act.

[136] Art 6 of the EU AI Act.

[137] Art 1 in the EUCFR.

[138] Arts 7 and 8 in the EUCFR.

[139] Art 21 in the EUCFR.

[140] Art 23 in the EUCFR.

[141] Art 11 in the EUCFR.

[142] Art 12 in the EUCFR.

5.5 AI and human dignity

Human dignity as a right and as a value is at the centre of international human rights law. As a self-standing right, it speaks to the autonomy, intrinsic self-worth and the protection of humanity in itself as its fundamental aims.[143] As a value, it is an aspirational standard in relation to other rights in any human rights instrument.[144] It reinforces the notion of the human being deserving respect in relation to other human beings, the human being as not being a tool to achieve a certain purpose and the human being flourishing in accordance with other human rights. In the *African Charter on Human Peoples' Rights* (hereafter, the "African Charter"), the right to human dignity is also informed by the notion that each person's own legal status in the world needs to be recognised and human dignity is to protected against exploitation and degradation in the form of slavery, slave trade, torture, cruel, inhuman or degrading punishment and treatment.[145]

When looking at human dignity in relation to AI, conversations abound about what the nature of human autonomy and agency entail especially in relation to developing and deploying AI-type technologies.[146] When the same conversations look at AI alone in the context of human dignity, several concerns have been raised about AI's ability to replace

[143] Nagy, Noémi (2023). "'Humanity's new frontier': Human rights implications of artificial intelligence and new technologies". In: *Hungarian Journal of Legal Studies* 64(2). 239.

[144] Teo, Sue Anne (2023). "Human dignity and AI: mapping the contours and utility of human dignity in addressing challenges by AI". In: *Law, Innovation and Technology*. Vol 15 (1), 243.

[145] Article 5 in the African Charter on Human and Peoples' Rights. Available at: https://au.int/sites/default/files/treaties/36390-treaty-0011_-_african_charter_on_human_and_peoples_rights_e.pdf (accessed 23 June 2024).

[146] Teo, Sue Anne (2023). "Human dignity and AI: mapping the contours and utility of human dignity in addressing challenges by AI", op. cit., 244.

human beings in terms of work, speed of performance and completing tasks. Moreover, greater fears have been expressed about AI being used to create autonomous weapons or weapons of a nature that can still be manipulated through human intervention.[147] This is on a more comprehensive scale when looking at AI's relation to human dignity, but when looking at persons themselves, the bias and discriminatory permutations that AI has been shown to demonstrate have the potential to infringe rather significantly on human dignity. For example, an AI system that administers welfare support for vulnerable individuals based on algorithmic categorisation potentially infringes not only on the privacy rights of individuals, but it also "objectifies" these individuals by singling them out through an embedded bias, which also leads to "instrumentalising" them.[148] Aizenberg and van den Hoven identify three particular categories of violations that relate to AI and a violation of human dignity as a right:

- *Humiliation:* being put in a state of helplessness, insignificance; losing autonomy over one's own representation.

- *Instrumentalisation:* treating an individual as exchangeable and merely a means to an end.

---

[147] Nagy, Noémi (2023). "'Humanity's new frontier': Human rights implications of artificial intelligence and new technologies". In: *Hungarian Journal of Legal Studies* 64(2). 239. Teo, Sue Anne (2023). "Human dignity and AI: mapping the contours and utility of human dignity in addressing challenges by AI". In: *Law, Innovation and Technology*. Vol 15 (1), 256.

[148] Teo, Sue Anne (2023). "Human dignity and AI: mapping the contours and utility of human dignity in addressing challenges by AI". In: *Law, Innovation and Technology.* Vol 15 (1), 256-7.

- *Rejection of one's gift:* making an individual superfluous, unacknowledging one's contribution, aspiration, and potential.[149]

Humiliation and rejection of one's gift can be revealed in instances where AI is relied upon to make hiring decisions or where AI has the potential to lay off entire workforces within a particular discipline. It is argued that humiliation also includes bias, discrimination and even defamation of persons or groups.

Instrumentalisation is revealed in the context where AI, based on its own biases and discriminatory understandings of people or a particular context, comes to a decision or formulates an output based on data associated with such people, thereby reducing the "human being" to numbers or binary to substantiate its thinking.

Suggested solutions to strengthening the relationship between human dignity and AI include ensuring that AI inculcates the life experiences of persons, which should include but is not limited to those experiences of marginalised or vulnerable groups who are most susceptible to risks and harm caused by AI systems.[150] One of the aspects of AI that is emphasised is a disembodiment of intelligence and humanity that is constructed based on data, while life experiences would then bring the physical human being and human condition to the forefront of AI in its structure and functioning.[151] The retention of human intervention in developing, deploying and using AI within various international instruments also emphasises the priority of the human being as being the driving force behind AI that is trustworthy, safe and ethical.

---

[149] Aizenberg, Evgeni and van den Hoven, Jeroen (2020). "Designing for human rights in AI". In: *Big Data & Society*. (sagepub.com) (accessed 30 May 2024), 5.
[150] Teo, Sue Anne (2023). "Human dignity and AI: mapping the contours and utility of human dignity in addressing challenges by AI". In: *Law, Innovation and Technology.* Vol 15 (1), 271.
[151] Ibid., 277-278.

In this sense, it is worth considering the importance of *Ubuntu*, the foundation of African philosophy and African moral theory, which is a collectivist concept in nature meaning "A person is a person through other persons". The concept emphasises community, connectedness to others and nature and the important role that the human being plays in the eyes of other human beings.[152] It is worth noting that at the heart of *Ubuntu* is the notion of respect for human dignity, respect, growth and potential in a collective context. In the context of AI, *Ubuntu* emphasises the importance of human beings by looking at their intelligence not only in the cognitive domain but also in other cerebral and corporeal processes.

Furthermore, *Ubuntu* posits that individual intelligence is not separate from the intelligence of other human beings or other forms of life. What this means in the context of AI is that believing in human beings completely being superseded by machines does not accord with *Ubuntu* but rather that working with AI must account for the human being working with it in relationship with other human beings, thereby removing any sense of helplessness or isolation that someone would feel with AI.[153] When looking at language, it has been noted that AI works within a conceptual framework that is informed by a particular programming language set and methodology. Language is also an element of *Ubuntu* and concerns arise in the context about what languages will inform AI, for example, the main language informing AI presently is English but the question arises as to how AI will possibly accommodate other languages,

---

[152] Van Norren, Dorine Eva (2023). "The ethics of artificial intelligence, UNESCO and the African Ubuntu Perspective." In*: Journal of Information, Communication and Ethics in Society*. 21(1). 116-118.

[153] Van Norren, Dorine Eva (2023). "The ethics of artificial intelligence, UNESCO and the African Ubuntu Perspective." In*: Journal of Information, Communication and Ethics in Society*. 21(1). 118-19.

like the African Nguni language in which *Ubuntu* finds its roots.[154] Concerns also arise about how AI can impact on the meaning of language as well as on the linguistic and semantic quality of a language like English and what would need to be considered is a way to ensure that language, an essential facet of the human condition, of one's human dignity and of one's connection to other human beings through *Ubuntu*, is not disregarded or denigrated to meaningless text by AI or is regarded as inferior for AI to learn and implement.[155]

What grounds Ubuntu as a principle is also further amplified within the *Resolution on the need to undertake a Study on human and people's rights and artificial intelligence (AI), robotics and other new and emerging technologies in Africa,* a resolution that was adopted by the African Commission on Human Peoples' Rights (hereafter "ACHPR") on 10 March 2021. The resolution urges all state parties to ensure that AI, robotics and other new and emerging technologies are compatible with rights and duties in the African Charter and in other regional as well as international human rights instruments, to uphold human dignity, privacy, equality, non-discrimination, inclusion, diversity, safety, fairness, transparency, accountability and economic development. In addition, the resolution urges member states to ensure that AI, robotics and other new and emerging technologies, especially those imported from other continents, are made fit for the African context, adjusted to Africa's needs and imbued with African values like Ubuntu, African norms and communitarian ethos.[156]

---

[154] Van Norren, Dorine Eva (2023), op. cit., 120.

[155] Ibid., 120.

[156] African Commission on Human and Peoples' Rights. Resolution on the need to undertake a Study on human and peoples' rights and artificial intelligence (AI), robotics and other new and emerging technologies in Africa (2021). Available at: https://achpr.au.int/en/adopted-resolutions/473-resolution-need-undertake-study-human-and-peoples-rights-and-art (accessed 23 June 2024).

5.6 AI and freedom of expression

Freedom of expression as a right is said to be an essential right for the healthy functioning of any democratic society and it includes not only the right to "express ideas and thoughts through any medium" but also the right to "receive and impart information and ideas." The right is however not absolute – it is subject to certain limitations in that any such expression is not done to "offend, shock or disturb" others within a democratic societal context.[157] Through expression in any "medium" comes the notion that this also includes expression via the internet and this type of expression is also to be afforded the same type of protection as is afforded to traditional free expression and free speech.[158]

There is intriguing discourse about the intersection between freedom of expression and AI, starting with the notion of algorithms used by AI serving as what is known as algorithmic speech and whether this type of speech is worthy of constitutional protection. Reference to algorithmic speech is considered in light of how an "algorithm" is defined (although this definition is vast depending on the context). It refers to the instructions that are used by AI to generate a particular output. There is not a great deal of jurisprudence that has explored this facet of AI as yet but it is relevant to consider in light of the potential bias and discrimination that may result from an AI platform's particular algorithmic structure and substrata.[159] Of course, a greater risk lies in AI learning its biases and discriminatory stances by means of human intervention.

With AI being able to learn based on human context and human input, questions arise as to whether outputs produced by AI can be protected by

---

[157] Sears, Alan M. (2020) "Algorithmic Speech and Freedom of Expression". In: *Vanderbilt Journal of Transnational Law*. 53(4), 1335 – 1336.

[158] Sears, Alan M. (2020) "Algorithmic Speech and Freedom of Expression". In: *Vanderbilt Journal of Transnational Law*. 53(4), 1336.

[159] "Algorithmic Speech and Freedom of Expression", Ibid. 1337.

the right to freedom of expression and speech. It is argued that outputs from an AI platform cannot be regarded as expressions or "self-expression" on the part of AI because AI can only generate text and images based on prompts and on what it is "trained to know" or what it learns by way of context.[160] This indicates that AI is only able to give us information that we seek, without necessarily reasoning about it or evaluating whether or not it is appropriate. Where the right to freedom of expression and free speech is relevant lies with the developers, deployers and users of AI who, through their ability to guide AI in using certain expressions and speech, may have this protected by right. At the heart of this notion is that AI, developer and deployers should ideally seek to ensure that AI is trained and programmed to provide reliable data on current events or a prompt provided by a user because choices made by such developers and deployers about the sources that they use to train their AI on will directly or indirectly impact on their AI's output.[161]

This also raises the question about who is liable in the event that AI provides defamatory outputs based on bias, inaccurate or fake information and discrimination. In some instances, literature points to developers and deployers of AI being responsible for defamation made in AI outputs where these output were actually embedded into the AI by the developer or deployer, but in other instances, there is argument for a prompter, user or writer of AI material who deliberately defames another, although through AI, being held liable in their personal capacity but this is difficult to prove.[162] In the context of AI and foundation models, "red

[160] Volokh, Eugene. Lemley, Mark A. and Henderson, P. (2023) "Freedom of Speech and AI Output". In: *Journal of Free Speech Law*. 3(2), 651.
[161] Volokh, Eugene. Lemley, Mark A. and Henderson, P. (2023) "Freedom of Speech and AI Output". In: *Journal of Free Speech Law*. 3(2), 652.
[162] Volokh, Eugene. (2023) "Large Libel Models? Liability for AI Output". In: *Journal of Free Speech Law*. 3(2), 495 – 497.

teaming" is an exercise that researchers on AI carry out to identify anything that would predispose a particular model to generate some types of harmful speech or other harmful content. This tends to be a hard technical task to perform but it is an ongoing attempt to ensure that harmful output and speech is prevented on any AI platform.[163] Moreover, this has become a complex domain to navigate within the realm of free expression and free speech in AI but, particularly in the American context, the criteria for consideration that have guided the determination of defamation through AI outputs focus on looking at whether, as is the case with online content, the output "materially contribut[es] to [the] alleged unlawfulness" of it and whether the output can be "reasonably perceived as a factual assertion."[164] This is finding effect in the real world where OpenAI LLC, the proprietor of Chat GPT, is engaged in a defamation suit involving a radio host in Georgia, Mark Walters, who alleged that ChatGPT made false claims about him embezzling money from a gun-rights organization and this has allegedly caused damage to his reputation. The matter is the first of its kind concerning freedom of expression and plausible defamation by AI. It appears that whatever the court decides in this matter, in dealing with a myriad of challenges before it, will certainly impact on freedom of expression, freedom of speech, the law of defamation and the responsibility of AI, its developers and deployers in future lawsuits.[165]

---

[163] Henderson, Peter. Hashimoto, Tatsunori and Lemley, Mark. (2023) "Where's the liability in harmful AI speech?". In: *Journal of Free Speech Law*. 593-594.

[164] Volokh, Eugene. (2023) "Large Libel Models? Liability for AI Output". In: *Journal of Free Speech Law*. 3(2), 497-98.

[165] Cahill, Rebecca. "Open AI Defamation Lawsuit: The first of its kind" (2023). Available at: https://lawreview.syr.edu/openai-defamation-lawsuit-the-first-of-its-kind/ (accessed 22 June 2024).

5.7 Recommendations for ensuring alignment between AI and human rights

At this stage, the EU AI Act suggests several mechanisms that can ensure alignment between AI and human rights, namely provisions concerning prohibited AI practices,[166] provisions that emphasise the importance of human oversight in relation to developing, deploying and using AI falling within the categories stipulated within the Act,[167] provisions concerning the requirements that need to be met by high-risk AI systems,[168] regulations surrounding data and privacy protection in the context of AI,[169] provisions regulating transparency and information to users,[170] provisions governing the requirement of conformity assessments for providers of high-risk AI systems,[171] and provisions governing the development of AI as an innovation within the confines of an AI regulatory sandbox.[172] These provisions amplify the importance of human beings, and they indicate a commitment to protecting human rights in their various forms in the quest to construct trustworthy and safe AI that can be used in an ethical manner.

An interesting point of departure to consider considering provisions such as those mentioned above concerns how to make AI intrinsically aware of human rights in the quest to align it with human rights. This then speaks to what and how AI must learn to understand what human rights are and why they are important. Aizenberg and van den Hoven posit an intriguing design methodology in this regard, known as "Design for

[166] Art 5 of the EU AI Act.
[167] Art 14 of the EU AI Act.
[168] Art 8 of the EU AI Act.
[169] Arts 10 and 70 of the EU AI Act.
[170] Arts 13 and 52 of the EU AI Act.
[171] Arts 19 and 43 of the EU AI Act.
[172] Art 54 of the EU AI Act.

Values". The term "Design for Values" entails taking moral and social values and making them context-orientated design requirements for technological systems, with the assistance of both direct and indirect shareholder feedback throughout the design process.[173] This involves a process of "value specification" by taking values and converting them into norms which refer to a system's capabilities to support the desired values. This becomes a socio-technical exercise, for example, if one takes the right to privacy as a value, this is converted into norms in a socio-technical way by embedding them into the system as follows, depending on context: informed consent, confidentiality and the right to erasure or to be forgotten.[174] These norms are then further socio-technically modified into another requirement, for example, informed consent is implemented as a positive opt-in (that is, something requiring the user's permission) and then this can be specified further. What is important to note is that these values, norms and how they are socio-technically designed are not deductive or hierarchical in nature and there is flexibility on how they can be formulated by stakeholders depending on the context.[175] This "Design for Values" approach follows a tripartite methodology in its preliminary research, when looking at what values are at stake, how they can be reduced to norms and what the socio-technical requirement will be:

---

[173] Aizenberg, Evgeni and van den Hoven, Jeroen (2020). "Designing for human rights in AI". In: *Big Data & Society*, 2020 (sagepub.com) (accessed 30 May 2024), 3.

[174] Aizenberg, Evgeni and van den Hoven, Jeroen (2020). "Designing for human rights in AI", ibid., 3.

[175] Aizenberg, Evgeni and van den Hoven, Jeroen (2020). "Designing for human rights in AI", ibid., 3.

- *Conceptual investigations:* identify the stakeholders and values implicated by the technological artifact in the considered context and define value specifications.
- *Empirical investigations:* explore stakeholders' needs, views, and experiences in relation to the technology and the values it implicates.
- *Technical investigations:* implement and evaluate technical solutions that support the values and norms elicited from the conceptual and empirical investigations.

By conducting research using this tripartite methodology and converting the data obtained into feasible values, norms, and socio-technical requirements for AI to follow, AI can learn to support values associated with certain human rights, where these values are reduced to norms that the AI platform must then consider as part of its creation, training, and implementation.

5.8 Conclusion

The relationship between AI and human rights has been depicted in two ways, by looking at AI as a catalyst or a conundrum, and by looking at the intersection between the two concepts with innovation of AI and the left of the intersection and risks posed by it to human rights on the left side of the intersection. It has been shown that AI has great potential to change the way that the world works but what needs to be considered is that, while AI can perform as the "human" in the digital world, there has to be a connection to how it works through the brains, intelligence and emotions of human beings in the physical world. This means that AI has to, in light of the risks that it poses, be informed by not only what makes us human but also by human rights to which all human beings are entitled. Various risks to human rights by AI have been canvassed in this contribution, with suggestions made on how to safeguard against them in

the context of human rights and human perspectives posited in a variety of international instruments. The relationship between AI and human rights is still developing, with several challenges that are appearing, but the importance of the human being cannot once again be overstated in ensuring that this relationship only grows to benefit all humankind and the global community at large when developing, deploying, or using AI.

6

SPIRITUAL DIMENSIONS OF AI IN SOCIETY

Corna Olivier, Johannes Cronje, South Africa

6.1 Introduction

Although[176] it might look strange to find a chapter on spiritual dimensions in a book on Artificial Intelligence (AI), one must remember that Alan Turing himself, one of the pioneers of AI, already foresaw a degree of spirituality in the future of the technology. Spirituality is regarded as a core human dimension and together with awareness, connection, insight, and purpose considered integral to understanding human flourishing (Dahl et al., 2020). The role of human spirituality is increasingly being explored in mental health and well-being infields of research such as psychology, psychiatry, and nursing. (Jaworski, 2015). The relationship between non-material aspects of nature and human well-being underscores the importance of spiritual dimensions in shaping identities, experiences, and capabilities (Rodrigues et al., 2021). Ethically balanced AI

[176] *Corna Olivier* is an educator and researcher specialising in psychology, IT, and robotics. Her current work explores Generative AI's applications in education and psychometrics. *Johannes Cronje* is a professor of Digital Teaching and Learning at the Cape Peninsula University of Technology, previously serving as Dean of Informatics and Design.

governance rules should therefore be developed with consideration of spiritual dimension of humanity.

This chapter deals with the relationship between spirituality and technology, tracing its historical roots, and examining its contemporary manifestations within the landscape of AI ethics and governance. Firstly, the chapter will explore historical attitudes towards technology and spirituality and highlight instances where spiritual beliefs have influenced technological advancements. Then follows, a brief overview of existing AI ethics frameworks and regulations, with a critical analysis of how spiritual dimensions are currently addressed or overlooked in AI governance efforts. The analysis will identify gaps in current approaches and introduce an exploration of spiritual dimensions in AI. The chapter will consider the ethical implications of AI development from a spiritual perspective and discuss strategies for integrating spiritual considerations into AI governance frameworks. Finally, ethical benchmarks will be formulated, informed by spiritual principles.

6.2 Historical perspectives on the intersection of spirituality and technology

Spirituality and religion are often intertwined (Alexander, 2020), and their relationship should be acknowledged when discussing historical perspectives. The influence of religion on technology is widespread. It may not be directly apparent, as sets of ideas, ideals and diverse ways of thinking are not immediately thought of as forms of technologies. Yet, it is ideas, ideals and diverse ways of thinking that inspire the development of technology. It is therefore clear that religious beliefs can influence societies to develop, or prevent them from developing, technology that can enhance or detract from their environment. Religions' role in influencing technology spreads from communication to warfare. Without religion, Gutenberg may not have built the printing press,

providing the cornerstone for mass media communications. Without religion, the Crusaders would not have had the need for lighter and slimmer battle armour. And without religion, the Chinese alchemists would have had no reason to try to find a potion for immortality, to live as gods, instead ending up accidently creating gunpowder, irreversibly changing the face of warfare and the course of human history (Altieri, 2016).

Over the last 100 to 150 years modern sciences have put great effort into distinguishing themselves from religious and moral inquiries. Such inquiries came to be perceived as non-scientific primarily due to their reliance on methods that lacked empirical or quantitative foundations (Harrison, 2016). The divide between science and religion, between spirituality and technology grew, and the modernist period, the first half of the twentieth century, is generally considered to be a period marked by rationality, secularity, and persistent atheism (Boyle, 2021).

However, the societal dynamic of reactions against established norms, the emergence of opposing ideologies, and the subsequent reconsideration and re-evaluation of those positions—all of which lead to renewed interest in spiritual or religious practices—is a recurring pattern in human history and contemporary society. This illustrates man's constant search for meaning, identity, and connection in the face of shifting social, cultural, and technological environments. (Karakaya, 2023; Plante, 2008). In 2001, the University of Copenhagen in Denmark hosted the first worldwide conference on religion and the Internet. *Religious Encounters in Digital Networks* brought together researchers from all around the world to discuss their studies of religion online (Campbell, 2005).

Then, in 2020 the world was driven online by the COVID-19 pandemic. This included religion and spiritual practices. There were Neo-Pagan online rituals, and the hashtag #Nous-SommesUnis, meaning '*Beyond our loneliness we are united on earth*' circulated by French Muslims. Then Pope Francis's Urbi et Orbi blessing praying for the end of

the Covid Pandemic was streamed from a dark, empty St Peters' square on 27 March 2020 (Evolvi, 2022). Then followed by the release of ChatGPT in November 2022.

Despite their differences, both events drew worldwide attention and sparked changes in how people and societies deal with difficulties and advances. Opinions on this new technology has been divided from the start (Xing et.al. ,2024). Spirituality is strongly associated with scepticism towards AI. The emergence of spirituality shows that spiritual ideas are increasingly influencing attitudes towards technology. It can thus be argued that a thorough study of the different worldview factors that influence attitudes towards technological development is essential. Recognising and addressing these worldview factors will allow stakeholders to better understand and respond to scepticism about developing technologies, ultimately enabling informed discussions and decision-making in society (Većkalov et.al., 2023).

6.3 Current landscape of AI ethics and governance

The rapid evolution of AI necessitated a corresponding pace in the development of AI ethics, with an increased emphasis on translating ethical principles into tangible norms and governance to ensure that artificial intelligence technologies are utilised responsibly and ethically. Existing AI ethics frameworks and legislation include a wide range of publications and guidelines designed to address the ethical implications of artificial intelligence technologies. These frameworks can be classified as regulatory frameworks, normative frameworks, applicative frameworks, and evaluative frameworks. (Alshehhi, Cheaitou & Rashid, 2022). Academic research has focused on developing ethical norms for AI in a variety of fields, including education, research, military, and healthcare. (Taddeo et al., 2021; Quinn & Coghlan, 2021).

UNESCO has also been actively involved in the ongoing discussion over AI ethics. In 2021, the organisation hosted an Intergovernmental Meeting of Experts to draft the Recommendation on the Ethics of Artificial Intelligence, emphasising the significance of ethical considerations in AI development. (Vică, Voinea & Uszkai, 2021). UNESCO's work in this field has been instrumental in supporting the ethical governance in robotics and AI systems. (Winfield & Jirotka, 2018).

UNESCO released a guidance document on Generative AI in education and research in September 2023. The document stresses the importance of prioritising human well-being over technological development, to ensure that technology is used to empower people. The document also emphasises on data privacy and age-appropriate usage, to protect vulnerable users and children in particular. According to the document effective Generative AI (GenAI) integration requires collaboration among stakeholders, such as educators, policymakers, and technologists, and needs teamwork in navigating this new terrain. It stresses that adaptive rules are essential to keep up with breakthroughs in GenAI and to maintain alignment with growing ethical considerations (Holmes & Miao, 2023).

The use of AI, and its potential for spiritual purposes, remains underexplored, because spirituality and technology are not usually linked. A study involving Christians, Hindus, Buddhists, a Muslim, and a Pagan, that investigated how people who practise a religion perceive the concept of communicating with a chatbot in a spiritual setting showed divided results. Positive feedback included praise for the capacity of the chatbot to deliver valuable information, while participants were sceptical about the subject knowledge required to provide valuable information to a user in need (Asante-Agyei, Xiao & Xiao, 2022).

To ensure a holistic and ethically grounded approach to technological development, although spirituality may not be associated with AI

naturally, it is important to recognise and integrate spiritual dimensions into AI governance frameworks. The need for practical methods and processes to ensure the ethical development and deployment of AI technologies is clear (Rees & Müller, 2022). In addition, structuring and implementing ethical and regulatory principles in AI technologies is essential to address societal concerns and build trust in these systems (Winfield & Jirotka, 2018; Svetlova, 2022).

6.4 An exploration of spiritual dimensions of AI

One cannot investigate the spiritual dimensions of AI without the insights of Alan Turing, who developed the "Turing Test" for artificial intelligence (Guo, 2015). As early as 1948 Turing speculated about "intelligent machinery", and asked the question, "Can computers think?". He proposed that computer would have passed the Turing test if an independent person behind a screen, after asking a series of questions, were unable to tell the response of the computer from that of a human. At the basis of his thinking was that it should be possible to imitate the human mind, so that computers could compete with humans in all intellectual fields. In his so-called "heretical theory" Turing proposes the possibility of a machine that not only thinks, but that can think independently. Guo (2015) defines spirituality as the "reconceptualization of the self" and shows how Turing himself reconceptualized his own being in response to developments of his understanding of technology.

From a posthuman perspective, Caldero Hernández (2021) explores the intersection of AI and spirituality. He stresses the need for a harmonious integration of technology and the humanities. He highlights the mutability of reality and shows how, even in scientific terms, that which is "real" is hard to define. He argues, for instance, that even the argument

that an object remains stationary unless acted upon by a moving force, is problematic, since we have no clear standard of what stationary means.

At the same time, he warns against reductionist interpretations that oversimplify human existence. What is required is an interdisciplinary approach that combines scientific rigor with humanistic values to deepen our understanding of complex phenomena. Collaboration between technology and the humanities may produce deeper insights into cultural, ethical, and human-centered design principles. Caldero Hernandez calls for empathy, inclusivity, and a deeper appreciation for the interconnectedness of human experiences. He underscores the importance of integrating diverse perspectives to address societal challenges and enhance our ethical engagement with technological advancements.

In combining arguments from the authors above it is possible to extract a number of dimensions of spirituality and AI:

The role of technology and humanities: There is a need for a harmonious vision that integrates technology and the humanities for the progress of knowledge and the well-being of humanity and nature. The gap needs to be bridged between data-driven fields and spiritual or humanistic studies to the benefit of both domains.

Artificial intelligence and spirituality: Artificial intelligence can aid in processing vast amounts of data to enhance understanding in spirituality, psychology, philosophy, and humanities. Furthermore, AI can help to separate genuine spiritual concepts from misleading ones, thus contributing to a deeper exploration of the spiritual realm.

Interdisciplinary synergy: There is a need for interdisciplinary approaches that combine technical advancements in AI with the conceptual and methodological needs of intangible disciplines such as philosophy, psychology, religion, and education. Artificial intelligence should be integrated into educational settings to foster a holistic understanding of human complexities and promote deeper intellectual engagement.

Unity of reality and knowledge: There is an interconnectedness between different dimensions of reality, and interdisciplinary research can serve to illuminate diverse aspects of existence. AI can play a crucial role in processing vast amounts of data to reveal convergences and divergences, facilitating a deeper exploration of human intangible aspects.

6.5 Proposals for integrating spiritual considerations into AI governance

With large language models increasingly being used to answer questions, and with their ability to generate accurate human-like answers, the expectation is created that such AI will increasingly be used to provide advice to people. In fact, the first "chatterbot", ELIZA, created in 1966 acted as a "therapist", asking open-ended questions and then presenting appropriate responses (Bassett, 2019).

If large language models can simulate independent thought, then it may well be that they could suffer from mental disorders (Ashrafian, 2017). In such a case three factors would need consideration: (1) Were the models unintentionally programmed to have mental disorders, and could these be remedied through corrective programming? (2) If the models possessed consciousness and free will, did they develop mental illness spontaneously, contrary to their original programming? (3) If artificial intelligences independently developed mental illness contrary to their programming, could this indicate an initial shift towards human-like consciousness, followed by the emergence of mental illness?

The answer to the first question lies in the inevitable existence of bias, stemming from "the nature of training data, model specifications, algorithmic constraints, product design, and policy decisions" (Ferrara, 2023, p. 1). Even though the models may originally be programmed to be "free" of bias, these biases will automatically appear, based on the nature of both the initial programmer and the users. In answer to the second question,

regardless of free will, the models can develop "mental illness" based on the collective prompts received over time.

If AI is becoming increasing human-like, then it behoves that IT governance should also consider the human factors associated with AI. The obvious first spiritual aspect lies in the perceived (and real) threat that AI could take one's job. The second lies in the improper use of AI for personal or institutional gain. The impact of AI on assessment is well known, with ChatGPT passing university level physics exams (Kortemeyer, 2023).

Aspects to consider when integrating spiritual considerations with AI governance include ethical norms and principles, posthuman design, inclusive stakeholder involvement, transparency and accountability, and impact assessment.

Ethical norms and principles: AI governance must integrate ethical norms that align with spiritual values, such as empathy, compassion, and the consideration of the broader consequences of AI actions. Although humans are currently able to display unique levels of empathy and emotional intelligence, AI can play a strong role in focused support that frees up humans to play a stronger role (Terblanche et al., 2022).

Posthuman design: At the human level we should ensure AI systems are designed to enhance human well-being and spiritual growth, promoting ethical interactions, and supporting the intrinsic dignity of individuals, while at the same time realize that AI is able both to promote and hinder spiritual well-being (He, 2022).

Inclusive stakeholder involvement: Diverse stakeholders, including spiritual leaders and ethicists, should be involved in the governance process to reflect a wide range of spiritual and moral perspectives. In the global South, for instance, the philosophy of Ubuntu, which can be summarized as "I am a person through other persons" (Van Norren, 2023) places an emphasis on social personhood and communal values rather

than transhumanism, capitalism, and individualistic privacy norms. When evaluating AI applications, Ubuntu encourages sharing and equitable distribution of benefits, consensus decision-making, and intercultural approaches that promote communal well-being and respect for diverse cultural perspectives (Van Norren, 2023). This stance may come into conflict with notions of AI for individual benefit.

Transparency and accountability: Transparency in AI decision-making processes should ensure that accountability mechanisms are in place to address any ethical or spiritual concerns that arise. It is necessary to consider bottom-up approaches alongside typical (Western) rule-based approaches to AI (Van Norren, 2023).

Impact assessment: Regularly assess the social and spiritual impacts of AI applications, ensuring they do not harm but rather benefit society and uphold spiritual values. Currently AI regulation is primarily aimed at protecting corporate interests. What is needed is a form of AI management that engages with political, economic, and social implications of these algorithmic systems. (Charlesworth, 2020).

6.6 Ethical benchmarks and recommendations

The ten areas of focus in the UNESCO Recommendations on the Ethics of Artificial Intelligence (UNESCO, 2021) are discussed below.

1. Transparency and *explainability*: Ensuring that AI systems are transparent in their operations and decisions and can be explained to users and stakeholders. Transparency in data processing regulations is crucial, yet its application to AI systems is challenging. Transparency entails delivering information and clarifying processes.

Nevertheless, context and human behaviour greatly shape our perception of transparency. Due to complex variables including human actions and behaviours, research on human-computer and human-robot

interactions does not clearly identify the factors that build trust in technology. Just providing information or explanations about AI systems may not build trust in them.

Transparency should be an interactive process between developers and users and focus on dialogue and the exchange of information to build trust. Future regulations should view transparency as a way of facilitating communication, sharing information, and building trust among AI stakeholders (Felzmann et al., 2019; Hois et al., 2019).

2. Non-discrimination and equity: To promote fairness and equality in the process of developing and deploying of AI technologies. Technological progress should be balanced with equitable distribution. The international community should therefore regulate the just development of AI under international law. New and specific regulations should be developed to address the novel challenges presented by AI. Key areas that need immediate attention are data sovereignty, data privacy, and data localization. These are essential for promoting the equitable development of AI.

Critical issues around freedom of expression and human rights arise from the use of AI for content moderation and censorship by governments and online platforms. The challenge for international lawyers and policymakers lies in the prevention of harmful content while still protecting individual rights to access information. For effective AI governance it is necessary to protect fundamental human rights, such as privacy, data security, equality, and non-discrimination. If AI technologies are grounded in these rights, they can deliver significant benefits to all humankind (Karim, 2023).

Table 1 on the next page shows various dimensions of fairness that must be integrated into AI development and deployment (Leslie et al., 2024):

Table 1 Dimensions of AI fairness (adapted from Leslie et al., 2024)

Category	Explanation
Data Fairness	Ensures AI systems are developed using diverse, accurate, and representative data to avoid biases.
Application Fairness	Prevents AI systems from creating or exacerbating inequalities, discrimination, or injustices, aligning with the values of those affected.
Model Design and Development Fairness	Avoids the use of discriminatory processes or structures in AI models, ensuring they do not perpetuate historical patterns of discrimination.
Metric-Based Fairness	Employs transparent and justifiable fairness metrics in AI systems, making them accessible to relevant stakeholders and impacted individuals.
System Implementation Fairness	Guarantees proper training for AI system users, ensuring they understand the system's limitations, strengths, and potential biases.
Ecosystem Fairness	Minimizes the impact of wider societal structures and institutions on AI innovation, and so prevents the reinforcement or amplification of discriminatory power dynamics and inequitable outcomes for marginalized or disadvantaged groups.

These principles of fairness are required to build trust and to ensure that AI technologies contribute positively to society (Karim, 2023; Leslie et al., 2024).

3. Respect for human autonomy: Upholding the autonomy and agency of individuals in interactions with AI systems. AI differs from other sociotechnical systems in that it can take autonomous decisions, which is a

threat to human agency (Van de Poel, 2020). Table 2 shows six spheres of analysis for Autonomy support required for the appropriate capture of contradictory and downstream effects (Calvo et al., 2020).

Table 2 Spheres of Experience to support Autonomy in AI (Adapted from Calvo et al., 2020)

Sphere	**Description**
Adoption	Refers to the experience of a technology prior to use, including the factors influencing a person's decision to use it. This sphere encompasses marketing strategies, internal motivations, social pressures, and volitional choices that lead to the adoption of a technology.
Interface	Involves a user's direct interaction with the software, such as navigation, buttons, and controls. At this level, a technology supports psychological needs by enhancing competence through ease-of-use and autonomy through task/goal support and meaningful options.
Task	Focuses on discrete activities facilitated by the technology, such as tracking steps in a fitness app or adding events in a calendar software. Each task can contribute to varying levels of need satisfaction, with some tasks feeling unnecessary or forced while others are perceived as useful and willingly performed.
Behavior	Encompasses the overarching goal-driven activity enabled or enhanced by the technology. It represents the culmination of various tasks contributing to a specific behavior, such as step-counting contributing to the behavior of exercise.
Technology	Refers to the impact of technology on the fulfilment of psychological needs within the user's direct experience. It influences autonomy and overall well-being,

	potentially affecting one's sense of thriving. This sphere considers the holistic impact of technology on life and psychological fulfilment.
Life	Captures the extent to which a technology influences the fulfilment of psychological needs within life overall, potentially affecting one's sense of thriving. It reflects how technology impacts autonomy and wellness at a broader level beyond individual tasks and behaviors.

4. Prevention of harm: Taking measures to prevent AI systems from causing harm to individuals or society. Policymakers and AI researchers should actively collaborate to investigate, prevent, and mitigate potential malicious uses of AI. Researchers and engineers in the field must recognise the dual-use nature of their work, allowing concerns related to misuse to influence research priorities and norms. They should also proactively engage with relevant actors when harmful applications are foreseeable. Best practices from research areas with more mature methods for addressing dual-use concerns, such as computer security, should be identified and imported where applicable to the case of AI. To ensure a comprehensive understanding of these challenges, it is essential to involve a diverse range of stakeholders and domain experts in discussions (Brundage et al., 2018).

5. Responsibility: Holding developers, users, and stakeholders accountable for the ethical use of AI technologies. Designing, developing, and maintaining AI systems while implementing responsible AI practices pose several key challenges (Deshpande & Sharp, 2022). AI system builders often lack training in ethical issues and may not have the opportunity to raise concerns within their organizations, leading to a focus on technical development rather than moral considerations. Organizations may shift the responsibility of AI system decisions onto the builders,

which can negatively impact those who have no say in non-technical decisions. Users of AI systems may lack the technical understanding necessary to comprehend how their choices influence AI behaviour, hindering responsible implementation. Organizational priorities such as financial performance, shareholder value, and competitive pressures can overshadow ethical considerations in AI system behaviour. The legal status of AI systems and the assignment of legal obligations remain uncertain, with current expectations placing responsibility on humans, though this may evolve with increasing AI autonomy. Identifying and engaging a wide range of stakeholders, as shown in Table 3, is crucial for ensuring responsible AI impacts, but coordinating these diverse interests can be challenging.

Table 3. AI Stakeholder categories (Adapted from Deshpande & Sharp, 2022)

Stakeholder Category	**Specific Stakeholders**
Individual Stakeholders	Users, developers, engineers, researchers, AI experts, HCI researchers and experts, non-users of AI systems, non-AI experts
Organisational Stakeholders	Technology companies, professional bodies (e.g., ACM, IEEE), learned societies (e.g., Japanese Society for Artificial Intelligence), research institutes (e.g., Future of Life Institute, Alan Turing Institute)
National/International Stakeholders	Nation states (EU, US, UK, China, Australia, etc.), regional/national legislative/regulatory agencies, OECD, UNICEF, etc.
AI System Builders	Those engaged in designing, developing, and maintaining AI systems

Addressing these challenges requires a multi-faceted approach that involves not only technical expertise but also a deep understanding of ethical considerations, legal implications, and effective stakeholder engagement strategies in the development and maintenance of responsible AI systems (Deshpande & Sharp, 2022).

6. Privacy and data governance: Safeguarding the privacy of individuals and establishing responsible data governance practices. Organizations need to consider several key factors to develop effective data governance strategies for AI. Firstly, clear organizational structures need to be established, with designated roles, such as Chief Data Officers or Chief Ethics Officers, to ensure accountability. A risk-based approach is essential to identify and address potential issues. Regular assessments should be conducted to manage the risks that come with data or algorithmic errors, biases, or discrimination. Mechanisms like public scrutiny and audits can promote transparency and accountability and contribute to responsible AI system implementation. System-level governance should be in place throughout the entire lifecycle of data and algorithms. An organizational culture needs to be created that values ethical and spiritual considerations. (Janssen et al., 2020).

7. Social benefit: Ensuring that AI technologies contribute to the well-being and advancement of society. Floridi et al. (2018) make 20 recommendations for societal benefit. These are divided into five action points: Assessment, Development, Legal, Auditing and solidarity, and described in Table 4, on the next page.

Table 4 Action points for social benefit (Adapted from Floridi et al. (2018)

Action Point	**Explanation**
Assessment	Evaluate the capacity of existing institutions to address mistakes and harms caused by AI systems.
Development	Create frameworks to enhance the explicability of AI systems for socially significant decisions.
Legal	Establish legal procedures and improve IT infrastructure for scrutinizing algorithmic decisions.
Auditing	Develop auditing mechanisms to identify unwanted consequences like unfair bias in AI systems.
Solidarity	Implement a solidarity mechanism, possibly in collaboration with the insurance sector, for severe risks in AI-intensive sectors.

8. Sustainability: Promoting the sustainable development and deployment of AI systems. Key development directions for AI sustainability include adopting a sustainability mindset, emphasizing environmental sustainability as a core design principle, and optimizing system and infrastructure efficiency. Machine learning practitioners and researchers should focus on AI data, experimentation, training algorithm efficiency, and telemetry for tracking the carbon footprint. System designers should rethink hardware design principles to minimize the environmental footprint of processors, memory, and storage technologies. Data centre operators should invest in carbon-efficient scheduling to neutralize the operational carbon footprint as electricity consumption rises. Research should also address the training efficiency of deep learning models to enhance overall efficiency and reduce environmental impact. Furthermore, it is crucial to consider environmental footprint characteristics throughout the

hardware and Machine Learning model development life cycle. (Wu et al., 2022).

9.Accountability: Establishing mechanisms for accountability and oversight in the use of AI technologies. Four accountability goals for AI are: (1) compliance, (2) report, (3) oversight, and (4) enforcement. These goals are prioritized by policy makers depending on the missions of AI governance and the reactive or proactive use of accountability (Novelli et al., 2023). Compliance is emphasized when the governance mission involves harmonizing values for the design, development, and deployment of AI technologies, fostering a movement towards values such as inclusivity, fairness, and transparency, that promote trust in AI. Reporting is crucial for providing sector-specific practical guidelines on AI, though it may be insufficient for broader governance missions such as setting the political agenda. Oversight is essential for ensuring adequate public debate and accurate regulation, especially for broad objectives such as establishing a political agenda. However, focusing solely on oversight may not suffice for narrower missions. Enforcement is emphasized in governance missions concentrating on fulfilling the anticipated advantages of AI technologies (Novelli et al., 2023). There are three key drivers of enforcement: Government Warning or Nudging, Anticipation or pre-emption of legislation, or self-interest (Gutierrez, 2022).

10. Inclusion: Ensuring that AI technologies are inclusive and accessible to all members of society. To ensure inclusivity in AI development we must address biases, promote diversity, invest in education, establish inclusive governance, and operationalize inclusiveness. Addressing biases involves the implementation of checks in AI system life cycles. The promotion of diversity in AI teams reduces biases and enhances stakeholder diversity. Investing in AI education in developing countries will bridge the gap between richer and poorer nations. Inclusive governance frameworks establish and enforce standards that promote inclusivity.

Operationalizing inclusiveness translates this principle into practice through clear guidelines (Belinchon et al., 2019).

6.7 Conclusion

This exploration of spiritual dimensions in AI society has shown a clear interplay between spirituality and technology, that transcends conventional boundaries and invites a deeper reflection on the ethical implications of artificial intelligence. Beginning with historical perspectives on the intersection of spirituality and technology, we have shown how religious beliefs have influenced technological advancements, and shaped societies and cultures across time.

The investigation of ethical benchmarks and recommendations for policymakers, technologists, and ethicists, has stressed the importance of ethical standards informed by spiritual principles. The guidelines for evaluating the ethical implications of AI technologies through a spiritual lens, call for a more conscientious approach to technological development—one that is grounded in values of empathy, interconnectedness, and consciousness.

It is important to encourage dialogue and collaboration between religious and spiritual communities, technologists, policymakers, and ethicists. A holistic and spiritually grounded approach to AI governance, has the potential to assist in addressing societal concerns, to build trust in AI systems, and to cultivate a more inclusive and ethically informed technological landscape.

References

Ashrafian, H. (2017). Can Artificial Intelligences Suffer from Mental Illness? A Philosophical Matter to Consider. Science and

Engineering Ethics, 23(2), 403–412. https://doi.org/10.1007/s11948-016-9783-0

Alexander, J. K. (2020). Introduction: The entanglement of technology and religion. History and Technology, 36(2), 165-186. https://doi.org/10.1080/07341512.2020.1814513

Alshehhi, K., Cheaitou, A., & Rashid, H. (2022). Adoption frameworks for artificial intelligence in the public sector: A systematic review of literature. Proceedings of the 3rd South American International Industrial Engineering and Operations Management, 919-929. https://doi.org/10.46254/SA03.20220211

Altieri, D. (2016). The Influence of Religion on Human Technologies. Swarthmore College, 6. Available at: http://fubini.swarthmore.edu/~ENVS2/dan/Essay2.html [Accessed 20 April 2024].

Asante-Agyei, C., Xiao, Y., & Xiao, L. (2022). Will you talk about god with a spirituality chatbot? an interview study. https://doi.org/10.24251/hicss.2022.399

Bassett, C. (2019). The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. AI and Society, 34(4), 803–812. https://doi.org/10.1007/s00146-018-0825-9

Belinchon, E., Bollmann, B., Cussins Newman, J., Jobin, A., & Nakonz, J. (2019). Towards an Inclusive Future in AI. A Global Participatory Process. A Global Participatory Process (October 2019). Belinchon, E., Bollmann, B., Cussins Newman, J., Jobin, A., Nakonz, J.

Boyle, K. R. (2021). "And the Word was God": rejection, consideration, and incorporation of spiritual motivations in modernist literature. Available at: http://hdl.handle.net/20.500.12648/1874 [Accessed 21 April 2024].

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., & Filar, B. (2018). The

malicious use of artificial intelligence: Forecasting, prevention, and mitigation. ArXiv Preprint ArXiv:1802.07228.

Calderero Hernández, J. F. (2021). Artificial intelligence and spirituality. International Journal of Interactive Multimedia and Artificial Intelligence, 7(1), 34–43. https://doi.org/10.9781/ijimai.2021.07.001

Campbell, H. (2005). Spiritualising the Internet. Uncovering discourses and narratives of religious Internet usage. Online–Heidelberg Journal of Religions on the Internet: Volume 01.1 Special Issue on Theory and Methodology. Available at: https://archiv.ub.uni-heidelberg.de/volltextserver/5824/1/Campbell4a.pdf [Accessed 21 April 2024].

Calvo, R. A., Peters, D., Vold, K., & Ryan, R. M. (2020). Supporting human autonomy in AI systems: A framework for ethical enquiry. Ethics of Digital Well-Being: A Multidisciplinary Approach, 31–54.

Charlesworth, A. (2020). Regulating algorithmic assemblages: looking beyond corporatist AI ethics. Data-Driven Personalisation in Markets, Politics and Law, Forthcoming.

Dahl, C. J., Wilson-Mendenhall, C. D., & Davidson, R. J. (2020). The plasticity of well-being: a training-based framework for the cultivation of human flourishing. Proceedings of the National Academy of Sciences, 117(51), 32197-32206. https://doi.org/10.1073/pnas.2014859117

Deshpande, A., & Sharp, H. (2022). Responsible ai systems: who are the stakeholders? Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 227–236.

Evolvi, G. (2022). Religion and the internet: Digital religion, (hyper) mediated spaces, and materiality. Zeitschrift Für Religion,

Gesellschaft Und Politik, 6(1), 9-25. https://doi.org/10.1007/s41682-021-00087-9

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data and Society, 6(1), 1–14. https://doi.org/10.1177/2053951719860542

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. ArXiv Preprint ArXiv:2304.03738.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., & Rossi, F. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28, 689–707.

Guo, T. (2015). Spirituality as reconceptualisation of the self: Alan Turing and his pioneering ideas on artificial intelligence. Culture and Religion, 16(3), 269–290. https://doi.org/10.1080/14755610.2015.1083457

Gutierrez, C. I. (2022). Identifying incentives for the enforcement of artificial intelligence soft law programs. IEEE Transactions on Artificial Intelligence.

Harrison, P. (2016). Peter Harrison's Territories of Science and Religion: A Symposium. Zygon, 51(3). Available at: https://www.zygonjournal.org/article/14330/galley/29037/download/ [Accessed 21 April 2021].

He, Y. (2022). Artificial intelligence and socioeconomic forces: transforming the landscape of religion. Humanities and Social Sciences Communications, 11:602(2024), 1–10. https://doi.org/10.1057/s41599-024-03137-8

Hois, J., Theofanou-Fuelbier, D., & Junk, A. J. (2019). How to achieve explainability and transparency in human AI interaction. HCI International 2019-Posters: 21st International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21, 177–183.

Holmes, W., & Miao, F. (2023). Guidance for generative AI in education and research. UNESCO Publishing. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000386693 [Accessed 21 April 2024].

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493.

Jaworski, R. (2015). Anthropocentric and theocentric spirituality as an object of psychological research. Journal for Perspectives of Economic Political and Social Integration, 21(1-2), 135-154. https://doi.org/10.2478/pepsi-2015-0006

Karakaya, M. (2023). The evolution of Western spirituality: From ancient Greece and Rome to contemporary practices. Uludağ Üniversitesi Fen-Edebiyat Fakültesi Sosyal Bilimler Dergisi. https://doi:10.21550/sosbilder.1256854

Karim, R. (2023). Toward an International Law of Just AI Development. Journal of East Asia & International Law, 16(2).

Kortemeyer, G. (2023). Could an Artificial-Intelligence agent pass an introductory physics course? http://arxiv.org/abs/2301.12127

Leslie, D., Rincon, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, S. J., Burr, C., Aitken, M., & Katell, M. (2024). AI fairness in practice. ArXiv Preprint ArXiv:2403.14636.

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society, 1–12.

Plante, T. G. (2008). What do the spiritual and religious traditions offer the practicing psychologist? Pastoral Psychology. https://doi:10.1007/s11089-008-0119-0

Quinn, T. and Coghlan, S. (2021). Readying medical students for medical ai: the need to embed ai ethics education.. https://doi.org/10.48550/arxiv.2109.02866

Rees, C. and Müller, B. (2022). All that glitters is not gold: trustworthy and ethical ai principles. Ai and Ethics, 3(4), 1241-1254. https://doi.org/10.1007/s43681-022-00232-x

Resseguier, A. (2023). Ai research ethics is in its infancy: the eu's ai act can make it a grown-up. Research Ethics, 20(2), 143-155. https://doi.org/10.1177/17470161231220946

Svetlova, E. (2022). Ai ethics and systemic risks in finance. Ai and Ethics, 2(4), 713-725. https://doi.org/10.1007/s43681-021-00129-1

Taddeo, M., McNeish, D., Blanchard, A., & Edgar, E. (2021). Ethical principles for artificial intelligence in national defence. Philosophy & Technology, 34(4), 1707-1729. https://doi.org/10.1007/s13347-021-00482-3

Terblanche, N., Molyn, J., de Haan, E., & Nilsson, V. O. (2022). Comparing artificial intelligence and human coaching goal attainment efficacy. PLoS ONE, 17(6 June), 1–17. https://doi.org/10.1371/journal.pone.0270255

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. Minds and Machines, 30(3), 385–409.

Van Norren, D. E. (2023). The ethics of artificial intelligence, UNESCO and the African Ubuntu perspective. Journal of Information, Communication and Ethics in Society, 21(1), 112–128.

Većkalov, B., van Stekelenburg, A., van Harreveld, F., & Rutjens, B. T. (2023). Who is skeptical about scientific innovation? Examining worldview predictors of artificial intelligence, nanotechnology, and human gene editing attitudes. Science Communication, 45(3), 337-366. https://doi.org/10.1177/10755470231184203

Vică, C., Voinea, C., & Uszkai, R. (2021). The emperor is naked. Információs Társadalom, 21(2), 83. https://doi.org/10.22503/inftars.xxi.2021.2.6

Winfield, A. and Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences, 376(2133), 20180085. https://doi.org/10.1098/rsta.2018.0085

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., & Bai, C. (2022). Sustainable ai: Environmental implications, challenges and opportunities. Proceedings of Machine Learning and Systems, 4, 795–813.

Xing, Y., Zhang, J. Z., Teng, G., & Zhou, X. (2024). Voices in the digital storm: Unraveling online polarization with ChatGPT. Technology in Society, 102534. https://doi.org/10.1016/j.techsoc.2024.102534

7

INTEGRATING DHARMIC PHILOSOPHICAL TENETS INTO GLOBAL AI GOVERNANCE

Ethical Framework and Regulatory Implications for Artificial Intelligence

Rohit Govind, India / USA

7.1 Introduction

This paper[177] explores the intersection of Dharmic philosophical principles and modern artificial intelligence (AI) regulation, presenting a unique fusion of ancient wisdom and contemporary technology. The analysis delineates a series of logical arguments that demonstrate how the ethical and philosophical doctrines from various Dharmic traditions can contribute to the formulation of AI governance. It highlights Buddhist (Shünyaväda and Vijñānavāda), Hindu (Charvaka, Nyāya, Pūrva-Mīmāṃsā, Sāṃkhya, Vaiśeṣika, Vedānta and Yoga), Jain and Sikh

[177] *Rohit Govind* is a graduate student pursuing a Master's degree in Applied Machine Intelligence at Northeastern University, Boston MA, United States. He has a professional background in technology strategy and innovation and has worked on numerous large-scale projects with global management consulting firms. He is associated with Dharma Alliance, a Geneva-based think-tank that seeks to mainstream Dharmic principles in global policies, norms and standards.

traditions, proposing their application on the establishment of laws governing AI development and application. This paper offers insights into the development of ethically robust AI systems, advocating for regulations that not only ensure the technical efficacy of AI but also its moral and ethical soundness.

7.2 Buddhist traditions and AI

7.2.1 Shünyaväda and AI Adaptability

Philosophical premise: Buddhist Shünyaväda teaches the doctrine of emptiness (śūnyatā), suggesting that phenomena lack inherent nature and are interdependently originated. This concept, along with Kṣaṇikavāda (the theory of momentariness), which posits that all phenomena are transient and exist only for a moment, resonates with chaos theory and emergent properties in complex systems, where outcomes are not linear but result from multiple interacting variables. In AI, this translates to using probabilistic models and adaptive learning techniques to ensure systems can adjust to new data and changing environments.

Scientific parallel: This concept resonates with chaos theory and emergent properties in complex systems, where outcomes are not linear but result from multiple interacting variables.

Law of emptiness and probabilistic decision-making in AI: "AI systems must be designed with the understanding that data and contexts are inherently empty of fixed meaning, requiring systems to employ probabilistic and adaptable decision-making processes that reflect this emptiness, ensuring flexibility and responsiveness to new information."

Potential risks in AI development: AI systems may inadvertently perpetuate suffering or unethical outcomes if they are designed without consideration of their broader impact on mental health and well-being,

particularly in sensitive applications like predictive policing or social credit systems.

Mitigating risks through Shünyavāda and Kṣaṇikavāda philosophy: Embracing Shünyavāda and Kṣaṇikavāda, AI developers can design algorithms that inherently question the absoluteness of their outputs, incorporating mechanisms that assess and communicate the uncertainty and probabilistic nature of AI predictions, thus enhancing transparency and trustworthiness. Continuous learning algorithms and real-time data processing ensure that AI systems remain current and effective in dynamic environments.

7.2.2 Vijñānavāda and AI cognition

Philosophical premise: Vijñānavāda, also known as Yogācāra, posits that reality is a construct of the mind, asserting that what we perceive as external objects are merely mental representations.

Scientific parallel: This philosophy aligns closely with the concepts in cognitive science and neuro-philosophy, which study how the brain constructs reality based on sensory inputs and internal mental processing. The idea resonates with virtual reality (VR) and augmented reality (AR) technologies, where user experiences are shaped entirely by artificial sensory information processed by the brain, creating a perceived reality that doesn't exist physically.

Law of consciousness-centric AI development: "Artificial Intelligence systems must be developed with an emphasis on enhancing human consciousness and ethical understanding. AI should support and extend human cognitive capacities without supplanting them, ensuring that AI operations are subordinate to human ethical considerations."

Potential risks in AI development: Deep learning systems, while powerful, are often opaque ('black boxes'), making it difficult to understand how decisions are made. This lack of transparency can lead to ethical and practical issues in implementation.

Mitigating risks through Vijñānavāda philosophy: The Vijñānavāda focus on a deep understanding of consciousness can inspire approaches to make AI systems more interpretable and transparent. Techniques like feature visualization and layer-wise relevance propagation can help demystify the operations of deep neural networks, aligning with the Vijñānavāda view of clarity in perception and knowledge.

7.3 Hindu traditions and AI

7.3.1 Charvaka philosophy and empirical foundations of AI

Philosophical premise: Charvaka's emphasis on direct perception and empirical evidence as the basis of knowledge mirrors the scientific method foundational in AI development, which relies on empirical data for algorithm training and validation. This approach ensures that training data is empirically validated, avoiding overfitting and enhancing fairness by relying on unbiased, empirically gathered data.

Scientific parallel: This mirrors the scientific method, foundational in AI development, which relies on empirical data for algorithm training and validation.

Law of empirical verification for AI systems: "All Artificial Intelligence systems shall undergo rigorous empirical testing and validation prior to deployment, with continuous real-time data verification mechanisms to ensure reliability, accuracy, and ethical integrity of AI outputs. This process must be transparent and documented to facilitate independent reviews and audits."

Potential risks in AI development: The primary risk associated with a purely empirical approach, akin to Materialism, is the over-reliance on data which might not represent unseen or future scenarios, leading to overfitting or poor generalization in AI systems.

Mitigating risks through Charvaka philosophy: Charvaka's insistence on direct perception can be used to enhance the robustness of AI algorithms by integrating real-time data validation and adaptive learning approaches. This can mitigate risks related to data insufficiency and bias, ensuring AI systems remain flexible and accurate even as external conditions change.

7.3.2 Nyāya and rational structuring in AI

Philosophical premise: Nyāya's methodical approach to knowledge through reasoning and debate mirrors the structured logical processes in computer science, particularly in developing algorithms that require a foundation of strong logical reasoning. This emphasizes the importance of algorithm transparency and explainable AI, where the decision-making processes are fully documented and understandable to users.

Scientific parallel: This mirrors the structured logical processes in computer science, particularly in developing algorithms that require a foundation of strong logical reasoning.

Law of logical transparency and accountability in AI: "AI systems, especially those used in critical decision-making, must operate on logically transparent algorithms, where the reasoning processes are fully documented and understandable to users. These systems must be accountable for their decisions through regular ethical audits."

Potential risks in AI development: Rule-based systems, while reliable in structured environments, can be inflexible or fail in scenarios that require nuanced understanding or adaptation, potentially leading to rigid and inappropriate applications of rules.

Mitigating risks through Nyāya philosophy: Nyāya's emphasis on logical reasoning and debate can be utilized to develop more flexible rule-based AI systems that incorporate exceptions and contextual variations into their decision-making processes. Techniques such as fuzzy logic, which allows for reasoning with degrees of truth rather than fixed

binaries, can embody the Nyāya commitment to detailed and contextual logical analysis.

7.3.3 Pūrva-Mīmāṃsā and ritualistic precision in AI

Philosophical premise: Pūrva-Mīmāṃsā emphasizes the performance of duties with ritualistic precision and adherence to sacred texts, akin to the programming of AI systems where algorithms must follow specified protocols meticulously. This mirrors principles in software engineering and robotics where exact execution of programmed instructions is critical.

Scientific parallel: This precision is akin to the programming of AI systems, where algorithms must follow specified protocols meticulously.

Law of dutiful adherence in AI operations: "AI systems must be programmed to adhere unwaveringly to predetermined ethical guidelines and operational protocols, reflecting the Pūrva-Mīmāṃsā emphasis on duty. Deviations from these protocols must be justified through a documented ethical review process."

Potential risks in AI development: A narrow focus on performing predefined functions can lead AI systems to operate in ways that are overly rigid or inappropriate in changing circumstances, potentially leading to ethical lapses or failures in dynamic environments.

Mitigating risks through Pūrva-Mīmāṃsā philosophy: The philosophical teachings of Pūrva-Mīmāṃsā encourage a rigorous adherence to duty while allowing for interpretation based on context. This can be applied to AI by designing systems that not only follow predefined rules but also incorporate adaptive learning capabilities to adjust their operations based on real-time feedback and changing conditions. Adaptive algorithms, such as those used in reinforcement learning, can ensure that AI systems remain flexible and context-aware, thus fulfilling their 'duties' effectively across varying scenarios.

7.3.4 Sāṃkhya and dualistic AI management

Philosophical premise: Sāṃkhya philosophy distinguishes between Purusha (consciousness) and Prakriti (matter), advocating a dualistic approach to understanding the universe. In AI, this reflects the dual structure of control systems: the algorithmic base (Prakriti) and the user interface or oversight mechanisms (Purusha). It aligns with cognitive theories that differentiate between conscious decision-making and automatic processes.

Scientific parallel: In AI, this reflects the dual structure of control systems: the algorithmic base (Prakriti) and the user interface or oversight mechanisms (Purusha).

Law of dualistic harmony in AI systems: "AI systems must be designed and regulated to maintain a balance between their operational functions (Prakriti) and the ethical oversight (Purusha). This dualistic approach ensures that AI systems perform efficiently while being rigorously monitored to uphold ethical standards."

Potential risks in AI development: One of the key risks in AI, particularly with autonomous systems, is the abdication of moral responsibility, where decisions are left entirely to algorithms without adequate human oversight, potentially leading to unethical outcomes.

Mitigating risks through Sāṃkhya philosophy: Implementing Sāṃkhya's philosophy in AI involves establishing clear boundaries and roles between AI systems (Prakriti) and human oversight (Purusha). By enforcing a model where ethical oversight by humans is mandatory, AI systems can be kept under check, ensuring that moral responsibility remains with humans. Techniques like human-in-the-loop (HITL) can be institutionalized, where critical decisions by AI are reviewed and can be overridden by human operators, ensuring ethical considerations are always prioritized.

7.3.5 Vaiśeṣika and atomistic AI construction

Philosophical premise: Vaiśeṣika posits that everything can be reduced to atoms, which combine in various ways to form more complex entities.

Scientific parallel: This atomistic approach is analogous to building AI systems using modular architectures, where smaller, independent models or 'atoms' of functionality are combined to create more sophisticated systems. This concept mirrors principles in both molecular biology and object-oriented programming.

Law of categorical integrity and data ethics in AI: "AI systems must adhere to strict categorization of data according to clearly defined and ethically sourced categories. These systems must maintain data integrity, ensuring that all AI actions are traceable and aligned with the highest standards of data ethics."

Potential risks in AI development: Atomistic approaches in AI could lead to reductionist thinking, where complex human behaviors and social interactions are overly simplified into data points, potentially missing the nuances and leading to dehumanized decision-making.

Mitigating risks through Vaiśeṣika philosophy: Applying Vaiśeṣika's principles, AI development can focus on synthesizing these atomistic insights into a coherent whole that respects and represents the complexity of human life. Techniques that ensure holistic data integration and multi-dimensional analysis can prevent the loss of critical information and ensure that AI systems provide nuanced and contextually aware outputs.

7.3.6 Vedānta and non-dualistic AI integration

Philosophical premise: Shankara Vedānta's Advaita (non-dualism) teaches that the ultimate reality is Brahman, transcending all apparent distinctions and dualities. This philosophy aligns with the concept of universality in physics and the notion in computational theory that complex

behaviors can emerge from simple, universal laws. This is reflected in general AI algorithms that adapt and function across various domains without needing domain-specific tuning.

Scientific parallel: This philosophy aligns with the concept of universality in physics and the notion in computational theory that complex behaviors can emerge from simple, universal laws.

Law of non-duality in AI governance: "AI systems must recognise the non-dual nature of existence by treating all data and user interactions with impartiality, devoid of biases. AI must be developed and deployed without discrimination, promoting equality and justice in every algorithmic decision, reflecting the Advaita philosophy of non-separation and equality."

Potential risks in AI development: The potential risk is that AI systems might not recognise the interconnectedness of all data and user interactions, leading to biased and unfair outcomes.

Mitigating risks through Vedānta philosophy: AI systems must be designed to treat all data and user interactions impartially, promoting equality and justice. Ensuring that AI systems are free from biases requires ongoing ethical audits and inclusive design practices.

7.3.7 Yoga and holistic AI development

Philosophical premise: Yoga emphasizes the unity of mind and body, advocating for balance, self-regulation, and harmony. These principles align with holistic approaches in cognitive science and systems biology, which consider entire systems rather than isolated parts. Yoga's focus on balance parallels homeostatic principles in biology, which maintain system stability.

Scientific parallel: These principles align with holistic approaches in cognitive science and systems biology, which consider entire systems rather than isolated parts.

Law of balance and self-regulation in AI: "Artificial Intelligence systems used in health and wellness must promote balance and support human self-regulation without leading to dependence. AI should enhance human capabilities while encouraging individuals' autonomy and well-being."

Potential risks in AI development: There is a risk that AI technologies might become too intrusive or controlling, overshadowing human autonomy and leading to over-dependence on technological solutions for health and well-being.

Mitigating risks through Yoga philosophy: Emphasizing Yoga's principles, AI technologies can be developed to support human efforts rather than replace them, ensuring that these systems are used to enhance but not control human life. Self-regulating AI systems that adapt to individual user needs and promote autonomy, such as adaptive learning environments that adjust to the learner's pace and style, embody Yoga's focus on personal balance and growth.

7.3.8 Jain traditions and non-violence in AI systems

Philosophical premise: Jain philosophy advocates for non-violence (Ahimsa) and considers multiple perspectives (Anekantavada), promoting ethical plurality.

Scientific parallel: These principles align with the development of AI systems that prioritize harm minimization and cultural inclusivity, crucial in global applications.

Law of non-violence and multiperspectivity in AI development: "Artificial Intelligence systems must be developed and operated under strict adherence to the principle of non-violence, ensuring no harm to humans or the environment. AI must incorporate and consider multiple perspectives and cultural contexts to avoid bias and promote fairness across diverse global populations."

Potential Risks in AI Development: A significant risk in AI, particularly in autonomous technologies, is the potential for harm, whether unintentional (through errors) or through a lack of comprehensive understanding of ethical implications across different cultural and social contexts.

Mitigating Risks Through Jain Philosophy: Incorporating Jain ethics, AI systems can be programmed to prioritize non-harm and evaluate decisions from multiple viewpoints, thereby reducing risks of harm and increasing the cultural and contextual sensitivity of AI decisions, which is critical in global applications.

7.3.9 Sikh traditions and community-centric AI development

Philosophical premise: Sikh philosophy emphasizes community service (*sewa*), equality before God (Ek Onkar), and justice, advocating a holistic approach that considers the welfare of the entire community. These principles can guide the development of AI systems that are designed to benefit entire communities, ensuring that the technology serves the public interest and promotes social welfare. AI could be used to optimize resource distribution, enhance public safety, and improve community health services, ensuring that these benefits are equitably distributed.

Scientific parallel: These principles align with the development of AI systems that prioritize harm minimization and cultural inclusivity, crucial in global applications.

Law of community-centric AI development: "All AI systems must be designed and operated to promote the welfare of the entire community, ensuring that they contribute positively to social justice, equity, and communal harmony."

Potential risks in AI development: AI technologies might become too intrusive or controlling, overshadowing human autonomy and leading to over-dependence on technological solutions for health and well-being.

Mitigating risks through Sikh philosophy: Emphasizing Sikh principles, AI technologies can be developed to support human efforts rather than replace them, ensuring that these systems are used to enhance but not control human life. Ensuring AI systems are inclusive and designed to serve the common good, promoting fairness and equity in their operations.

7.4 Conclusion

The nascent field of AI regulation and policy requires a diverse and inclusive approach to ensure that this fundamental technology is bias-free, safe, transparent, and beneficial for all humans. Integrating philosophical insights from traditions around the world, particularly Dharmic philosophies, can significantly contribute to developing ethical AI governance. Dharmic philosophies, with their deep and relevant ethical discourses, offer valuable frameworks for ensuring that AI systems are not only technically proficient but also morally and ethically sound. By drawing on these ancient teachings and recognizing their logical arguments, robust AI regulations could be formulated that promote human welfare, fairness, and justice in the development and deployment of AI technologies.

Selected Bibliography

Wallach, W., Franklin, S., & Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. *Journal of Artificial Intelligence Research*.

Cervantes, J. A., Rodríguez, L. F., López, S., & Ramos, F. F. (2013). A Biologically Inspired Computational Model of Moral Decision Making for Autonomous Agents. *Journal of Autonomous Agents and Multi-Agent Systems*.

Cervantes, J. A. et al. (2013). Cognitive Process of Moral Decision-Making for Autonomous Agents. *Cognitive Systems Research*.

Wallach, W. (2010). Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making. *Ethics and Information Technology*.

Kadkhoda, M., & Jahani, H. (2012). Problem-solving Capacities of Spiritual Intelligence for Artificial Intelligence. *Journal of Applied Research in Intellectual Disabilities*.

Guo, T. (2015). Spirituality as Reconceptualization of the Self: Alan Turing and His Pioneering Ideas on Artificial Intelligence. *Journal of Religious Studies*.

Vaughan, F. (2002). What is Spiritual Intelligence? *Journal of Humanistic Psychology*.

Tan, C. (2020). Digital Confucius? Exploring the Implications of Artificial Intelligence in Spiritual Education. *Journal of Educational Technology*.

Graves, M. (2021). Emergent Models for Moral AI Spirituality. *Journal of AI and Society*.

Prasad, H. (2023). The Buddhist Pramāṇa-Epistemology, Logic, and Language: with Reference to Vasubandhu, Dignāga, and Dharmakīrti. *Journal of Indian Philosophy*.

Islam, M. (2022). Anekāntavāda and Its Relevance: A Philosophical Analysis in Jaina Viewpoint. *Journal of Jaina Studies*.

Balcerowicz, P. (2016). Siddhasena Mahāmati and Akalaṅka Bhaṭṭa: A Revolution in Jaina Epistemology. *Journal of the American Oriental Society*.

Divakaran, A., Ramya, S., & Srinivasan, R. (2022). Broadening AI Ethics Narratives: An Indic Art View. *Journal of Ethics in AI*.

8

REGULATIONS ON THE USE OF GENERATIVE AI IN RESEARCH

Karla Sofia Molina Araniva, Switzerland / El Salvador

8.1 Introduction

In recent years[178], generative Artificial Intelligence (AI) has emerged as a powerful tool with diverse applications across various fields, including scientific research and academic writing. With the ability to generate original content, such as text, audio, and visual data, based on learned patterns and styles from existing datasets, generative AI has significantly impacted the way researchers approach tasks such as text generation, summarisation, and data analysis, among others. The availability of Large Language Models (LLMs) has further facilitated these processes, enabling researchers to enhance their productivity and efficiency in producing written content, especially across multiple languages.

However, the increasing use of AI in research also presents challenges related to research integrity, authorship, plagiarism, and data protection, thereby sparking important discussions within the research community, asking for common regulations and recommendations on the use of these tools. Considering this, in this exploratory essay, I will explore the

[178] *Karla Sofia Molina Araniva* is a Philosophy and International Relations graduate, from the Geneva Graduate Institute.

existing regulations regarding the utilisation of generative AI within the realm of research, paying particular attention to the cases of the European Union and the Americas (comprising the United States and Latin America and the Caribbean). This exploration will provide insight into the distinct regulatory dynamics present in these regions, emphasising the shared concerns that shape their approaches.

8.2 How is generative AI relevant to research?

In general terms, generative AI encompasses a series of techniques that leverage datasets to create new and original content, including text, audio, visual, and graphical data, based on the patterns and styles learned from existing data. Generative AI tools can generate diverse outputs responding to a range of simple or complex prompts, commands, or inquiries. Amid its widespread applicability, the LLM constitutes a specific class of AI algorithm tailored for tasks associated with natural language, encompassing functions like text generation, summarisation, and predictive analytics. Upon receiving a prompt, the LLM can produce contextually relevant outputs in various formats, including essays, narratives, extensive or concise responses, and conversational exchanges.[179]

In the context of research, generative AI tools have been increasingly utilised for a variety of purposes, from assisting in the production of human-like texts, language translation, and even in summarising and synthesising large volumes of data. These tools have shown promise in facilitating literature review processes and enhancing the ability to gather insights from diverse sources such as papers and social media.

[179]Andrés Castellanos-Gómez, "Good Practices for Scientific Article Writing with ChatGPT and Other Artificial Intelligence Language Models," *Nanomanufacturing* 3 (2023): 135–138, https://doi.org/10.3390/nanomanufacturing3020009.

In particular, the availability of publicly accessible AI and LLM models for research applications has shaped tasks associated with drafting reports, grammar checking, data analysis and visualisation. The use of these models enables researchers to accelerate the writing process (traditionally considered to be laborious and time intensive) especially for individuals who are not native speakers, as it assists them in generating written content in multiple languages.[180]

Despite the positive advantages that AI and LLM models present for research, the emergence of tools capable of generating human-like texts has raised concerns about the implications for research integrity and the credibility of the findings. AI and LLM models have the ability to create realistic looking but entirely synthetic (fake) data, which may be mistaken for genuine one. Since it is difficult to trace the origins of the data or to replicate the results, this could lead to the inclusion of false information in research studies, thereby compromising the credibility of the findings and the overall research integrity. The latter raises concerns about the potential for spreading misinformation and fake data, or the misrepresentation of data and results, ultimately affecting the credibility and reliability of the research.[181]

Debates around authorship[182], plagiarism and content validity have also been present among the research community, which are confronted with the dilemma of whether AI-generated content warrants the same

---

[180] Jingshan Huang and Ming Tan, "The role of ChatGPT in scientific communication: writing better scientific review articles," *Am J Cancer Res* 13, no. 4 (2023): 1148–1154, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10164801/.

[181] Tuba Livberber and Süheyla Ayvaz, "The impact of Artificial Intelligence in academia: Views of Turkish academics on ChatGPT," *Heliyon* 9, no.9 (2023) https://doi.org/10.1016/j.heliyon.2023.e19688.

[182] Manuela Marescotti, "To ChatGPT or not to ChatGPT: the use of artificial intelligence in writing scientific papers," *Brain Communications* 5, no.6 (2023) https://doi.org/10.1093/braincomms/fcad266.

recognition as human-authored work. This raises the question of how to appropriately credit AI algorithms for their contribution, leading to discussions on whether they should be listed as co-authors or acknowledged in a different manner. Additionally, there is a growing concern regarding plagiarism and the validation of research findings when considering integrating AI-generated content in publications. While AI tools can produce content that closely resembles human writing, they often do not provide clear sources for the information they generate. As a result, confirming the originality and accuracy of AI-generated content can be challenging.

On another note, legitimate worries about data protection have been increasing in the broader research community, mainly when sensitive data or internal organisational information is used to train AI models and generate new content. If sensitive data becomes part of the broader online environment, this could lead to potential security and privacy issues, particularly when personally identifiable information is at stake. Unauthorised access to such data could have significant consequences for research institutions and individual researchers, increasing the potential for misuse and exploitation.

8.3 Regulations on the use of generative AI in research: The cases of the European Union and the Americas

8.3.1. The case of the European Union: The European Commission's guidelines on the responsible use of generative AI in research

These guidelines focus on one particular area of AI used in the research process, namely generative Artificial Intelligence. There is an important step to prevent misuse and ensure that generative AI plays a positive role in improving research practices. One of the goals of these guidelines is that the scientific community uses this technology in a responsible manner.

European Commission

Issue: In light of the growing integration of generative AI tools in the research process, numerous European research institutions have established their individual guidelines for the use of these tools. The different regulations have created a complex landscape, making it challenging to decide which ones to follow under specific contexts. To address this issue, the European Commission, working through the ERA Forum Stakeholders, has taken the initiative to establish common guidelines for the integration of generative AI in research. These are intended to provide direction for funding bodies, research organisations, and researchers in both public and private settings. The primary objective is to encourage the responsible use of generative AI in the scientific community, ensure transparency in its application, and address concerns related to data integrity, privacy, and ethics.[183]

Response: The European Commission has outlined four principles behind the guidelines on the use of generative AI in research:

1. **Reliability**: This involves ensuring the quality of research through the design, methodology, analysis, and resource usage while also addressing issues related to bias and inaccuracies.
2. **Honesty**: Researchers are expected to develop, carry out, review, report, and communicate research transparently and impartially, including disclosing the use of generative AI.
3. **Respect**: Responsible use of generative AI should consider its limitations, environmental impact, societal effects, and proper information management, privacy, and intellectual property rights.

[183] European Commission, Living guidelines on the responsible use of generative AI in research, (Research and Innovation, 2024).

4. **Accountability**: Researchers are accountable for their research from idea to publication, including its management, broader societal impacts, and oversight.

Considering these principles, the European Commission emphasises the key role of researchers in ensuring the reliability of the content generated by or with the support of AI tools. This means that they are urged to maintain a critical approach by being aware of generative AI's limitations and potential biases and being cautioned against using fabricated material created by the tool, that can be considered false or incorrect. Regarding research organisations, the European Commission suggests that they should promote, guide, and support on the use of generative AI by providing trainings, especially on verifying output, maintaining privacy, addressing biases, and protecting intellectual property rights. Additionally, they should ensure compliance with ethical and legal requirements such as EU data protection rules and intellectual property rights.[184]

On another note, transparency in the use of generative AI is a cornerstone of these guidelines. Researchers are not only encouraged, but also expected to detail the tools used in their research processes and provide information on how they were utilised and affected the research. This transparency fosters a sense of accountability and trustworthiness. Additionally, particular attention must be paid to issues related to privacy, confidentiality, and intellectual property rights when sharing sensitive information with AI platforms, to prevent unauthorised or unintended reuse of that data. On this topic, funding bodies are encouraged to design funding instruments that are open, receptive, and supportive of ethical and responsible generative AI use. Additionally, they should

[184] Jamie Berryhill, Kévin Kok Heang, Rob Clogher and Keegan McBride, *Hello, World: Artificial intelligence and its use in the public sector*, (OECD, 2019).

request transparency from applicants on their usage of generative AI and actively monitor its use.

The guidelines consider researchers as ultimately responsible for the content generated by or with the support of generative AI tools. This means that these tools are not considered authors or co-authors, since authorship implies human agency and responsibility[185]. Although the use of generative AI should be acknowledged, but these tools cannot be attributed as author. Researchers must also be mindful of potential plagiarism when utilising outputs from generative AI and give proper citation to the original creators when necessary. The ultimate goal of these guidelines is to ensure the responsible and transparent use of generative AI in research. They are not binding, and researchers, institutions and funding bodies are encouraged to adapt them to their specific contexts while maintaining proportionality.

8.3.2 The case of the Americas

8.3.2.1 The United States and the Blueprint regulation

The White House Office of Science and Technology Policy has identified five principles that should guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence. The Blueprint for an AI Bill of Rights is a guide for a society that protects all people from these threats—and uses technologies in ways that reinforce our highest values.

The White House Office of Science and Technology Policy

Issue: The US government recognises that the potential misuse of unproven or inadequately functioning technology accompanies the deployment of technologies to address diverse issues. This misuse can result in significant harm, whether due to malfunctioning technology or

[185] European Commission, Living guidelines on the responsible use of generative AI in research, (Research and Innovation, 2024).

intentional safety transgressions. Besides, automated systems may introduce irrelevant historical data into unrelated decision-making processes.[186] Intended or unintended uses of technology can also lead to unforeseen harmful consequences. For these reasons, the US government has developed a regulatory framework based on the voluntary principle in the context of ongoing global discussions on AI regulation and policy-making. This framework, while not yet specific to the use of generative IA in scientific research, as in the case of the European Union, is part of a larger conversation on using AI in all social domains.

Response: One significant soft law mechanism is the *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*, released by the White House Office of Science and Technology Policy in October 2022. The Blueprint intends to support the development of policies and practices to protect civil rights through designing, implementing, and governing AI systems. It focuses on national values and incorporates various resources and best practices to enhance transparency and reliability in these systems and decision-making processes.[187] The Blueprint outlines five fundamental principles[188]:

1. **Ensuring safe and effective systems:** Emphasises the importance of safeguarding against the improper or irrelevant use of data in creating, advancing, and implementing automated systems, as well as shielding against the amplified negative effects of its reuse.

[186] The White House Washington, Blueprint for an AI bill of rights making automated systems work for the American People, (n.p., 2022).
[187] Le Cheng and Xiuli Liu, "Unravelling Power of the Unseen: Towards an Interdisciplinary Synthesis of Generative AI Regulation," *Int. J. Digit. Law Gov* 1, no. 1 (2024): 29–51, https://doi.org/10.1515/ijdlg-2024-0008.
[188] The White House Washington, Blueprint for an AI bill of rights making automated systems work for the American People, (n.p., 2022).

2. **Protecting against algorithmic discrimination**: Those creating, advancing, and implementing automated systems must consistently and actively take steps to safeguard individuals and communities from algorithmic discrimination and to create and use systems fairly.
3. **Safeguarding data privacy**: Being safeguarded from harmful data practices with built-in protections and having control over the data usage.
4. **Providing notice and explanation:** It is important to recognise that an automated system is in use and understand how and why it influences outcomes.
5. **Offering alternative options:** Provide the option to choose a human alternative over automated systems when suitable. Suitability should be assessed based on reasonable expectations in a specific context, with an emphasis on promoting broad accessibility and safeguarding the public from particularly harmful effects.

Although these principles lack regulatory enforcement and the Blueprint does not constitute a legislative and executable act for rights protection, it represents a forward-looking governance blueprint based on future aspirations.[189] In addition, this document has been the basis for the creation of guidelines on the use of generative AI in research by several universities and research institutions in the United States. This includes, for example, the case of the Arizona State University and Harvard University, which have created guidelines targeting the use of generative AI tools in research process, highlighting their commitment to promoting the responsible experimentation of those tools. However, considering factors such as information security and data privacy, compliance, copyright, and academic integrity when utilising generative AI in research studies.

---

[189] European Parliament, *At a Glance. United States approach to artificial intelligence*, (European Parliament, 2024).

Among the guidelines of these institutions[190], it is suggested that researchers should avoid inputting confidential data into publicly accessible generative AI tools to prevent the exposure of sensitive information to unauthorised parties. This includes refraining from using sensitive and internal data in externally sourced tools. On another note, the guidelines also aim to address concerns about the accuracy, bias, and potential fabrication of AI-generated content, as well as the possibility of it containing copyrighted material.

AI-generated content often draws from existing sources, leading to worries about plagiarism and intellectual property rights. In this sense, the different guidelines propose verifying the AI-generated content using trustworthy sources and stress the role of the researchers in having the ultimate responsibility for any content they publish that includes AI-generated material.

8.3.2.2 Latin America and the Caribbean: the ECLAC regulation initiatives

Issue: The regulatory landscape for AI in Latin America and the Caribbean is still in the early stages of development. As of now, there are no comprehensive region-wide guidelines specifically targeting AI, as in the

[190] Arizona State University, "Guidelines for Use of Artificial Intelligence (AI) in Research," n.d., https://researchintegrity.asu.edu/export-controls-and-security/artificial-intelligence.Cornell University, "Guidelines for Artificial Intelligence," n.d., https://it.cornell.edu/ai/ai-guidelines. Harvard University, "Initial guidelines for the use of Generative AI tools at Harvard," n.d., https://huit.harvard.edu/ai/guidelines. Standford University, "Generative AI Policy Guidance," February 16, 2023, https://communitystandards.stanford.edu/generative-ai-policy-guidance. The University of Iowa, "Using Artificial Intelligence (AI) Tools in Research. Guidance on using AI tools in research," n.d., https://its.uiowa.edu/support/article/127731. U.S. National Science Foundation, "Artificial Intelligence," n.d., https://new.nsf.gov/focus-areas/artificial-intelligence.

case of the European Union, and there is even less about the use of generative AI for research-related purposes. Nevertheless, there is a growing concern among the States about the potential and existing risks associated with AI.[191] As a result, individual countries have started to address AI's ethical, legal, and social implications, and regional initiatives, mainly led by the ECLAC are also emerging. Furthermore, collaborative ties are being initiated between the European Union and Latin America and the Caribbean, with the aim of boosting digital cooperation through concrete actions that promote the development of digital infrastructure and the convergence of policies and regulations that guarantee the safeguarding of human rights online.

Responses: A study conducted by the National Center for Artificial Intelligence (CENIA) shows that among the outstanding experiences in AI governance, the cases of Brazil, Argentina and Chile stand out for having a series of projects carried out and legal initiatives underway, along with the adoption of international instruments on AI. Argentina has developed a series of state recommendations for using AI to protect fundamental rights, preventing or diminishing risks and biases. Thus, for example, in the field of research and education, it is recommended that AI actors do everything reasonably possible to avoid reinforcing or perpetuating discriminatory and biased outcomes. In this sense, it is highlighted the important role of humans in verifying the AI outputs and application because an AI system cannot replace human beings' ultimate responsibility and accountability.

In Argentina and Brazil, bills have been introduced before the Congress of Deputies and the State to regulate intellectual property

[191] Centro Nacional de Inteligencia Artificial (CENIA) *Índice Latinoamericano de Inteligencia Artificial* (CENIA, 2023).

concerning productions made or assisted by AI[192]. In both countries, the aim is to ensure that all content produced or assisted by AI is not protected by intellectual property law and that protection is only granted to the part of the work resulting from human intellect. On the other hand, the Economic Commission for Latin America and the Caribbean (ECLAC) has taken significant steps in creating a regional agenda on the use of AI by launching the first Latin American Artificial Intelligence Index in 2023. This index aims to measure and assess the development and impact of AI in the region and provide valuable data and insights to help policymakers, researchers, and industry stakeholders understand the current state of AI in the region.

The 2023 Digital Agenda for Latin America and the Caribbean (eLAC) is a strategy that proposes the use of digital technologies as instruments for sustainable development. Its mission is to promote the development of the digital ecosystem in Latin America and the Caribbean through a process of regional integration and cooperation, strengthening digital policies that promote knowledge, inclusion and equity, innovation, and environmental sustainability. The Agenda addresses topics such as fostering innovation, ensuring privacy and data protection, and promoting digital skills and literacy to ensure that the benefits of AI are widely accessible.

Also in 2023, ECLAC, in coordination with the European Union, has promoted the Digital Alliance[193], an initiative that will seek to accelerate the digital transformation process in Latin America and the Caribbean

[192] Sophia Alphin Arévalo, "América Latina: Novedades en materia de inteligencia artificial en relación con la propiedad intellectual," May 13, 2024, https://institutoautor.org/america-latina-novedades-en-materia-de-inteligencia-artificial/.

[193] CEPAL, "América Latina y el Caribe y la Unión Europea refuerzan sus lazos de colaboración con el lanzamiento de la Alianza Digital EU-LAC," March 15, 2023, https://www.cepal.org/es/notas/america-latina-caribe-la-union-europea-refuerzan-sus-lazos-colaboracion-lanzamiento-la-alianza.

through capacity building to improve the design and implementation of digital policies, regulations and governance mechanisms. It will also seek to consolidate a space for permanent dialogue between the two regions, strengthen the institutional capacities of the countries and support the productive transformation based on digitalisation and innovation in Latin America and the Caribbean.

8.4 Conclusion

In the cases reviewed, the regions present different and contrasting progress in their AI frameworks, which makes it difficult to make a comprehensive comparison among them, regarding regulations around artificial intelligence. However, it is worth to make notice that there are shared concerns about the risks of engaging with AI systems.

These concerns include the potential for bias and inaccuracies in research due to the use of AI tools, the need for transparency in the application of AI tools and systems, and the importance of addressing issues related to data integrity, privacy, and ethics. All three regions emphasise the responsible use of AI and highlight the key role of human individuals in having the ultimate responsibility for AI-generated outputs. Additionally, there is a shared focus on encouraging transparency, providing proper information management, and safeguarding intellectual property rights when using AI.

9

GOVERNANCE OF SUPPLY CHAIN RISKS: THE ROLE OF ARTIFICIAL INTELLIGENCE (AI)

Kushani De Silva, Sri Lanka

9.1 Introduction

Today's global supply chains[194] are more concerned with international policies, strategic controls, and compliance mechanisms to protect supply chain partnerships. The concern of supply chain governance plays an important role beyond supply chain management which is more of a coordinated operation where human factors in technology integration in management and governance become pivotal. For example, while leveraging Artificial Intelligence (AI) technology is advantageous, the human factor remains crucial to success. Humans contribute unique qualities to business processes and operations, such as ethical judgment, creativity, emotional intelligence, and accountability. AI-based machines are faster and more accurate. However, humans bring intuition, emotion, and

[194] *Kushani da Silva*, PhD, is the founder and CEO of Gloir.K, a non-profit focused on education in disaster risk reduction, disarmament, counter-terrorism, and environmental conservation. She also advises UN agencies on Disaster Risk Reduction and collaborates with CRDF Global and CEEDASIA. Additionally, she is a member of the Centre for Women's Research.

cultural sense, adding greater significance to the workforce. Therefore, AI and human employees lead to more effective decision-making and innovation in various industries.

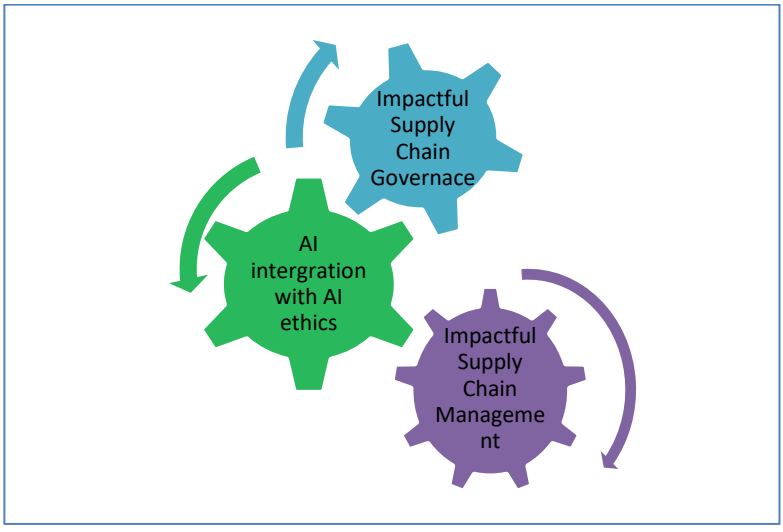


Figure 1: The relationship among supply chain governance, supply chain management, and AI ethics integration. Source: (De Silva, 2024).

The relationship among supply chain governance, supply chain management, and AI ethics integration can be illustrated in Fig 1. It reflects AI ethics as the driving force for impactful supply chain management through impactful supply chain governance.[195,196]

[195] Supply Chain Governance Is Essential to the Ethical Business. December 14, 2021. Retrieved from: https://www.logility.com/blog/supply-chain-governance-vs-management/

[196] De Silva, K. (2022), Today's Critical Need for Disarmament and Non-proliferation of Weapons of Mass Destruction (WMD); Treaties, Emerging tech threats, and gender inequality/inequity inclusive terrorism in perspective. Gloir.K. Sri Lanka. ISBN 978-624-6323-00-4.

AI should only enhance human qualities, not replace them. For example, AI has created opportunities in healthcare diagnoses. Nevertheless, human skills are essential for prescribing, and executing treatments through rational thinking, and emotional intelligence. Human connection can be enhanced through AI built-in social media however, content creation with ethics and cultural values needs human skills.

AI increases labour efficiencies by automating tasks which is inadequate. Supply chain governance in automation needs human skills that respect AI ethics to prevent human rights violations, misuse of technology, and compliance monitoring of end-users and end-use of dual-use goods and technologies. The enforcement of ethical limits for AI use is essential for reducing/preventing human rights violations, and climate degradation by exploiting supply chain processes. Also, to prevent the dissemination of disinformation that creates social risks.[197]

Another aspect of AI integration is its ability to process data to resemble intelligent behaviour reshaping the world order, human interactions, and our living styles. Therefore, AI ethics are required to protect humanity, individuals, societies, the environment, human rights, and human dignity.

Strategic supply chain governance with AI includes mitigating risks, controlling trade, decreasing operating costs by reducing labour costs, and improving relationships with suppliers, manufacturers, and distributors to improve satisfaction.[198] AI facilitates the online delivery of

---

[197] Ramos, G (2024), Assistant Director-General for Social and Human Sciences of UNESCO, Social and Human Sciences Sector at a glance. Retrieved from: https://www.unesco.org/en/social-humansciences/about#:~:text=Gabriela%20Ramos%20is%20the%20Assistant,social%20inclusion%20and%20gender%20equality

[198] De Silva, K. and Perera, R. (2024). Strategic Trade Control of Transshipments: Know Your Customer Based Best Practices for Counterproliferation, *Strategic*

required stocks by tracking shipments and analyzing large amounts of weather and conflict-related data within a short period of time. Further, analyze sales patterns for optimizing inventory levels, reducing product stockouts, and getting fresh food to shoppers faster. AI can find the most effective way (efficient and cost-effective) to transport goods analyzing the number of trucks needed, and fuel consumption by integrating Uninterruptible Power Supply (UPS), traffic plans, and fuel consumption with accurate and updated data.

Overall objective: Given the explained background, the objective of this article is to reveal the role of AI ethics in the contexts of AI applications yesterday, today, and tomorrow, the change of world order with AI creating supply chain governance risks, predicting unpredictable business contexts, and the risk of AI challenging global security norms. The next sections illustrate the role of AI ethics yesterday, today, and tomorrow.

9.2 The role of AI ethics yesterday, today, tomorrow

AI is a dynamic technology hard to explain due to its complexity and uncertainties. However, risk reduction experts can also set logical and moral grounds to discuss AI in the eyes of reducing global risks.

The concept of Artificial Intelligence was proposed by British mathematician Alan Turing (1912-54) in 1950 while leading the first symposium on Artificial Intelligence held in 1956 in Dartmouth, New Hampshire, USA. As a result, the international community of scholars officially recognised AI as a science. Therefore, as for any science discipline, AI too should have AI ethics compulsorily.

Trade Review. Volume 10, Issue 11, Winter/Spring 2024. Retrieved from; https://strategictraderesearch.org/wp-content/uploads/2024/02/Strategic-Trade-Control-of-Transshipments-Know-Your-Customer-Based-Best-Practices-for-Counterproliferation.pdf

AI is reshaping the world, and the US has long been and remains, the global leader in AI.[199] Kai-Fu Lee is well-known AI pioneer in China once highlighted that most people have no idea and many people have the wrong idea about the potential of AI. In terms of AI, potential governments' ability to use data to control people operating as business entities is one example. It is better to highlight that super-fast computer chips, all online world data, and deep learning computer programming make this task easier. The ultimate result will be a new world order of power which means changing structures and international rules. This calls for immediate attention to respecting AI ethics by creating AI ethical frameworks as appropriate.

After seventy years since the first theoretical approaches, AI is widely used in an increasing number of production and human life surpassing the human brain's performance in specialized fields. This is a result of deep research capabilities translating into commercial value. For example, China introduces new business models even for smaller businesses like shared bicycle services, and ordering small packaging online (E.g. Daraz) to be prominent in digital markets. Chinese companies are more willing to adopt the top-down approach in public data sharing, personal information sharing, and customizing AI Apps and other systems to local needs. For example, the changes made to Alibaba Cloud in the Middle East would nearly impossible with Amazon Cloud Drive. Chinese companies prioritize developing new methods with phenomenal access to data sets.

For example, other countries pioneering AI might not have databases of small package purchases/customers/perceptions that the Chinese companies have through their localized and customized new small-scale integrated business models.

---

[199] Lee, K.F. (2018) AI Superpowers: China, Silicon Valley, and the New World Order

AI helps analysis of large data sets within a short period to reach conclusions, but human oversight is essential for ensuring ethical use of such data. Further, for AI to be used ethical and user-friendly it should be easily understandable even to a non-expert. This is called the *explainability* of the AI model output process. There is not a "yes/no" approach to AI output generation as it operates on a spectrum, and the user should be well aware of the significant consequences of different AI model outcomes. Further, ensuring bad actors' attempts to find minor variations in the AI to produce undesired outputs and hackers attempting the get access when the AI model is learning from an online data source (for example Microsoft's AI chatbot incident in 2016) highlights the need for ensuring AI security as a key AI ethical consideration.[200]

In terms of big data sets Tencent and Alibaba have some of the most powerful datasets everywhere. For example, these two companies have mobile payments 100 times higher than PayPal and 10 times higher than Master or Visa. This highlights the need for an overarching data protection and privacy framework, such as the General Data Protection Regulation (GDPR) in the EU, to be introduced in China.[201] Selling data is a criminal offense in China.

UK, Germany, and France have strong AI research track records. However, Europe needs to fully utilize the required venture capital entrepreneur's ecosystem as the US and China. Europe is strong in hardware and telecoms, but less strong than the US or China in developing

[200] Valori, G. E. (2023). Artificial Intelligence and the New World Order (I), articles published on July 8, 2023. Retrieved from: https://moderndiplomacy.eu/2023/07/08/artificial-intelligence-and-the-new-world-order-i/

[201] EU GDPR (European Union General Data Protection Law. (2024). What is GDPR, the EU's new data protection law? Retrieved from: https://gdpr.eu/what-is-gdpr/

consumer internet companies, social media companies, and huge mobile application companies.

It is important to protect the privacy of people which is a main concern of the UK and Europe which could also be achieved by developing an innovative application that respects idealism. Exploring the connectivity of big data through international collaboration is needed within an AI ethical framework. For example, China's hyper-engaged users and big data sets can be linked to the UK's cutting-edge capabilities through the AI ethics framework respected by the parties engaged. This is due to past evidence that combined capabilities can advance innovations. For example, the UK's top AI company DeepMind, beat the world champions at the ancient Chinese board game Go using its AlphaGo programme in 2016. Likewise, the possibilities could be explored how social media video app such as Tik-Tok could be modified for creating enriched content with a positive societal message by integrating AI ethics.

9.3 Change of world order with AI creating supply chain governance risks

Responsible supply chain governance requires a governance strategy, transparency demanded by the consumers, internal and external collaboration with stakeholders, and the right technology to track and audit suppliers/end-users with real-time data and reporting mechanisms. Contracts, standards, mechanisms for reporting, and social bonds are some elements of governance mechanisms. Further, collaborative planning, communicating plans with policy guidelines, mitigation planning based on emerging risk assessment, and developing compliance metrics with required competency-based training are essential for supply chain governance to be successful.

Supply chain management through good governance and human skills is a strategy to prevent misuse and avoid illicit end users. With the

integration of AI improving transparency, legitimate last-mile delivery, tailored solutions, disruption reduction, and agile procurement have been achieved. For example, enhancing transparency throughout the supply chain and reducing last-mile delivery risks by ensuring the legitimate use of goods and services. However, it is worth noting that AI ethics plays an important role in this. Rodriguez et al. (2022)[202] highlighted that Explainable AI (XAI) is required for exploring the combination of human and machine capabilities, as well as ethical implications to ensure the appropriate use of AI.[203]

AI integration has been prominent in the Agriculture sector in India. A study conducted by Nayal et al. (2022) revealed that process AI helps real-time information sharing, storing, and analyzing, enhancing the supply chain management process overall. For example, AI integration facilitates the immediate acquisition of information, allowing for proactive decision-making and risk mitigation by immediate acquisition of information.[204]

---

[202] M. Rodriguez-Sampaio, M. Rincón, S. Valladares-Rodríguez, M. Bachiller-Mayoral (2022). - Explainable artificial intelligence to detect breast cancer: A qualitative case-based visual interpretability approach. International Work-Conference on the Interplay between Natural and Artificial Computation, Springer, Cham (2022), pp. 557-566.

[203] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, Francisco Herrera (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion, Volume 99, 2023, 101805, ISSN 1566-2535, Retrieved from: https://www.sciencedirect.com/science/article/pii/S1566253523001148

[204] Nayal, K., Kumar, S., Raut, R.D., Queiroz, M.M., Priyadarshinee, P. and Narkhede, B.E. (2022), "Supply chain firm performance in circular economy and digital era to achieve sustainable development goals", Business Strategy and the Environment, Vol. 31 No. 3, pp. 1058-1073.

9.4 Predict unpredictable business contexts

AI-powered Google's search engine enables Facebook to target advertising and support Alexa and Siri to perform their jobs. Predictive policing, risks of self-driving cars, and autonomous weapons that kill without human intervention are some examples of risks of using AI without ethical limits to predict unpredictable business contexts.[205]

AI revolutionizes supply chain management through predictive analytics. For example, AI inbuilt demand planning software has become popular as it can predict and estimate costs, production capabilities, and ability to deliver during extreme weather events, economic crises, etc. which enables organizations to adopt new supply chain practices. Supply chain optimizing software also helps monitor product quality, balance inventory levels, and identify fuel-efficient delivery routes efficiently. AI can accurately predict future demand by analysing data, allowing companies to optimize inventory levels, streamline supply chain processes, and reduce the risk of stockouts or overstocking. According to research, 37% of businesses, including supply chain companies, already see the benefits of AI solutions. AI will also contribute $15.7 trillion to the global economy by 2030.[206]

AI analysis based on events and development trends that influence international relations/collaborations is a common practice. Therefore, the need to use AI complying with AI ethics in managing risks and challenges in international collaborations has become pivotal.

---

[205] Mark, C. (2020). AI Ethics, MIT Press, ISBN: 9780262538190.

[206] Patel, V. (2023). How Artificial Intelligence Is Revolutionizing Supply Chain Management. Article published on July 28, 2023. Retrieved from: https://www.computer.org/publications/tech-news/trends/ai-revolutionizing-supply-chain

9.5. AI challenging the global security norms

AI has become an influential factor in the macroeconomy and military defence spheres. The emergence of non-state sector actors' nonpeaceful use of AI applications for supply chain attacks is becoming a key consideration in supply chain security management, creating threats to global trade and peace and supply chain governance.

AI has the potential to challenge existing international laws and ethics challenging governance and security norms. Therefore, AI ethics built on global security norms become pivotal. It is not the internet and cyber science that threaten the security norms but the human skills to misuse AI. This further justifies the use and application of AI needs to be part of countries' ethical frameworks and penalties should be imposed for violating such AI ethics.

It is worth noting some global predictions by individuals. For example, Alvin Toffler in his book The Third Wave published in the United States in 1980 and Italy in 1987 predicts that the future world will be tensed with nuclear weapon risks and creating economic and ecological collapse. Thus, supply chain governance through AI integration needs to be done with adequate AI ethics enforcement to avoid future world power struggles/conflicts.

The emergence of nuclear technology changed the modern world's political landscape and further strengthened the power structure of the major powers shaped at the end of World War II. For example, permanent members of the Security Council of the United Nations facilitated a series of international rules, such as the use of nuclear energy for peaceful purposes; the nuclear States' commitment to the non-proliferation of nuclear weapons, and the non-nuclear States' access to peaceful nuclear technology. At the same time, they enacted a series of international agreements such as the Treaty on the Non-Proliferation of Nuclear Weapons, the Comprehensive Nuclear-Test-Ban Treaty, the UN Negotiating

Mechanism for Nuclear Disarmament, the Global Nuclear Security Summit, and the Southeast Asian nuclear-weapon-free Zone.

The use of AI without ethical considerations can lead to a situation where dual-use of AI (peaceful and non-peaceful purposes) can be misused challenging the balance of international power. The potential can be assessed to a greater extent by considering the AI capabilities in AI ethical frameworks illustrated in Fig 2.

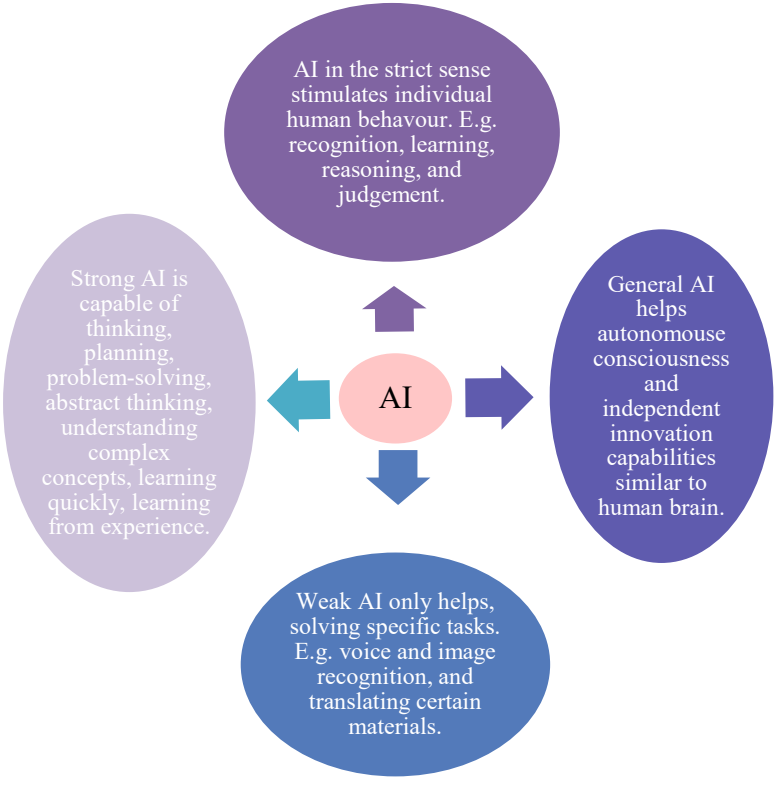


Figure 2: AI capabilities that need to be considered in AI ethical frameworks. Source: (De Silva, 2024)

The continuous evolution of AI technology itself combined with a rigorous system of prevention of misuse through ethics and laws, are essential for balancing power struggles in global trade and geopolitics making international collaborations more transparent and consented.

9.6 Conclusion

This article revealed the role of AI ethics in the contexts of AI applications yesterday, today, and tomorrow, the change of world order with AI creating supply chain governance risks, predicting unpredictable business contexts, and the risk of AI challenging global security norms. As such leveraging AI technology is identified as advantageous, but the human factor remains crucial to success. Humans contribute unique qualities to business processes and operations, such as ethical judgment, creativity, emotional intelligence, and accountability.

Further, brings intuition, emotion, and cultural sense, adding greater significance to the workforce. The relationship among supply chain governance, supply chain management, and AI ethics integration can be illustrated in Fig 1 and it reflects AI ethics as the driving force for impactful supply chain management through impactful supply chain governance.

AI has become an influential factor in the macroeconomy and military defence spheres. The emergence of non-state sector actors' nonpeaceful use of AI applications for supply chain attacks is becoming a key consideration in supply chain security management, creating threats to global trade and peace and supply chain governance. AI helps analysis of large data sets within a short period to reach conclusions, but human oversight is essential for ensuring ethical use of such data. Further, for AI to be used in an ethical and user-friendly way it should be easily

understandable even to a non-expert. This is called the *explainability* of the AI model output process.

Even if there is not a "yes/no" approach to AI output generation as it operates on a spectrum, the user should be aware of the significant consequences of different AI model outcomes. The continuous evolution of AI technology itself combined with a rigorous system of prevention of misuse through ethics and laws, are essential for balancing power struggles in global trade and geopolitics making international collaborations transparent.

10

CHALLENGING OBLIGATIONS: INVESTIGATING THE 'AI LITERACY' MANDATE FOR HUMAN OVERSIGHT

Amanda Horzyk, Poland / United Kingdom

10.1 Introduction

The imminent publication[207] of AI literacy provisions in EU's Official Journal raises a new mandate for deployers to ensure "to their best extent" an adequate level of AI literacy of those assigned with human oversight. These provisions, introduced late in the discussion of the proposed AI Act, follow a precautionary approach. Arguably they depart from the initially anticipated "product liability" framework focusing on the obligations of system providers. Deployers must now ensure that persons assigned with human oversight duties have the necessary competence to fulfil the task, particularly an adequate level of AI literacy, training and authority. To ensure that they have the skills, knowledge and

[207] *Amanda Horzyk* is a researcher in technology law, particularly in areas of Internet Law, Artificial Intelligence Regulation, AI Ethics and Virtual Reality. Her current research focuses on bridging the legal and technical perspectives in developing and evaluating leading solutions to complex issues presented by recent developments in AI, Virtual Reality, and associated Law.

understanding necessary for the informed deployment and use of AI systems and their output; given their opportunities, risks and possible resultant harms.

Research Question: Given the importance of effective Human Oversight in mitigating, preventing or eliminating the risks of High-Risk AI-Systems ('HRAIS') on individual's health, safety and fundamental rights – is the 'AI Literacy' mandate sufficient to ensure that Human Overseers can fulfil their role?

The publication of the Artificial Intelligence Act ('AI Act') in the EU's Official Journal raises a new mandate for deployers to ensure "to their best extent" an adequate level of AI-literacy of those assigned with human oversight.[208] These provisions, introduced late in the policy discussion, follow a precautionary approach. Deployers must now ensure that persons assigned with human oversight duties have the necessary competence to fulfil the task, particularly an adequate level of AI-literacy, training, and authority.[209] Given the importance of effective Human Oversight in mitigating, preventing or eliminating HRAIS risks on individuals' health, safety and fundamental rights, the paper asks whether the 'AI-literacy' mandate is sufficient to ensure that Human Overseers can fulfil their role as per Article 14 of the EU AI Act. It is argued that the 'AI-Literacy' mandate has several implications; but, whether they are sufficient to ensure effective human oversight will depend on the notions of (i) necessity and (ii) quality of AI-literacy. Moreover, it will rely on

[208] European Parliament and Committees, 'CORRIGENDUM to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts ('Corrigendum')' (19 April 2024) cor01 art 4.

[209] Ibid. rec 91.

whether AI-literacy as demanded by the act, is met with effective implementation by relevant players in the supply chain.

After providing a necessary context of the policy understanding of 'AI-literacy', and its origin in the AI Act (10.2), the paper will dissect the two elements of the 'AI-literacy' mandate (10.3). The strongest opinions and main developments of the AI-literacy provisions, will clarify what the 'quality' and 'necessity' elements encapsulate in the final version of the Act. The evolution will expose the implicit intentions of the participants to the regulatory formation; constituting the driving force behind the final mandate. The paper will then discuss what is the role of Human Overseers as per Article 14 (10.4) to arrive at what 'literacy' is called for (10.5). It is revealed how the binary delineation of AI-literacy is attached to the context of use (intended application domain, supply chain actors' responsibilities, and surrounding organizational and technical infrastructure), which is far from static. The work concludes that AI-literacy alone is not sufficient to ensure that Human Overseers can fulfil their role under Art.14, as it constitutes just one pillar supporting the required 'competence' enabling and empowering them to fulfil the law's propositum. Meanwhile, how this pillar holds, will be relative to AI-systems' context and implementation.

10.2 AI-literacy in policy

The promotion of digital, data and artificial intelligence literacy is perceived as a vital initiative in the reform of European Education; and

was already undertaken by several member states,[210] but not all.[211] It involves educational movements that are essential in fostering a responsible and ethical AI future; as we continue to monitor AI's potential adverse effects, misuse, and research developments.[212] It is considered a necessary part of a public dialogue, civic participation and training, so that "all members of society can make informed decisions about their use of AI-systems and be protected from undue influence".[213] As most, remain unaware of the input AI-systems use or what consequences it may have on them.[214] It is to promote high-quality training and education empowering all to become active participants in the society, economy and democracy.[215] Accordingly, sectors, including education,[216] culture and

[210] Education European, Agency Culture Executive and Union Publications Office of the European, 'Informatics education at school in Europe : Eurydice report' (2022) 38; e.g. Canada, Finland, Australia see Sutherland Eric, Anderson Brian and Oecd, *Collective action for responsible AI in health*, 2024) 28; Science Canada. Innovation, issuing body Economic Development Canada and Canada Government of, 'Learning together for responsible artificial intelligence : report of the public awareness working group.: Iu173-38/2022E-PDF Publications - Canada.ca' (2022).

[211] Education European, Agency Culture Executive and Union Publications Office of the European, 'AI report - Publications Office of the EU' (2024), 52.

[212] Ibid. 76.

[213] UNESCO Recommendation on the Ethics of Artificial Intelligence ('UNESCO Recommendations on AI Ethics') (Paris, 2022), SHS/BIO/PI/2021/1 para 44.

[214] Science, Committee Technology and U. K. Parliament Select Committee Publications, 'GAI0084 - Governance of artificial intelligence (AI)' (18 July 2022) 6.

[215] Corrigendum (n 1), rec 56.

[216] E.g. Schoolnet European, *Data4Learning: ChatGPT and the role of AI in assessment*, 2023) 6; as reiterated by Council's recommendation, see Union Council of the European, 'Council Recommendation on the key enabling factors for successful digital education and training' (2023) 22, para 3(f).

creatives[217] plead for EU States to develop appropriate AI-literacy frameworks to guide professional development and transition into this algorithmic era.

While AI-literacy has long been included in computational literacy education, it is a relatively new concept in policy discourses attempting to formally conceptualize it.[218] Hence, the shift towards developing AI-literacy, in most relevant policy documents, is still largely targeted at developing educational programs that build citizens' *general* understanding of AI's impact, capabilities and limitations, usefulness and questionable applications, and how it may be used for public good.[219] Most policy documents to date rely on Duri Long's classification of AI-literacy competencies, including (1) knowing and understanding basic AI, and AI-applications' functions (2) to apply and use AI concepts in various scenarios (3) create and evaluate AI and (4) converse in AI ethics, among others.[220]

---

[217] Plenary European Parliament, 'European Parliament resolution of 21 November 2023 with recommendations to the Commission on an EU framework for the social and professional situation of artists and workers in the cultural and creative sectors (2023/2051(INL))' (2023) para 48, 61.

[218] Davy Tsz Kit Ng and others, 'Conceptualizing AI literacy: An exploratory review' (2021) 2 Computers and Education: Artificial Intelligence 100041; Whitehouse.gov, 'NSTC: Building Computational Literacy Through STEM Education: A Guide for Federal Agencies and Stakeholders' (2023) 23.

[219] UNESCO, *K-12 AI curricula: a mapping of government-endorsed AI curricula*, 2022) 11.

[220] Duri Long and Brian Magerko, 'What is AI literacy? Competencies and design considerations' (2020) Proceedings of the 2020 CHI conference on human factors in computing systems 1, cited by e.g. Whitehouse.gov, *NSTC: Building Computational Literacy Through STEM Education: A Guide for Federal Agencies and Stakeholders* 23; OECD, *OECD Employment Outlook 2023*, 2023) 179; UNESCO, *K-12 AI curricula: a mapping of government-endorsed AI curricula*, 15.

A broader understanding of 'literacy' that empowers the wider public, across all levels and countries.[221]

However, UNESCO's approach to 'AI-literacy'[222] aligns more closely with that required by the AI Act (sufficient to enable appropriate Human Oversight, see 10.3). It highlights 'AI-literacy' as involving a *level of competence*: skill, knowledge, understanding and value orientation (*inter alia,* principles of ethics on the responsible and trustworthy design, development and deployment of AI).[223] It encompasses not only data literacy (i.e. knowing how algorithms collect, analyze, manipulate and clean data), but also algorithm literacy (i.e. understanding how AI identifies patterns and makes connections within datasets), understanding key concepts of AI techniques (e.g. Neural Networks, Reinforcement Learning) and AI technologies (e.g. Computer Vision, Natural Language Processing).[224]

The latter formulation is also associated with 'value-orientation', a constructive approach to developing *general* understanding of ethical challenges raised by AI.[225] For instance, controversies arising from ceding control to AI-systems and the need for human oversight.[226] Where an application is considered alongside a *societal need;* in this instance, the attribution of legal and ethical responsibility over AI-systems decision-making (or AI-assisted decision-making), and the need of operationalizing redress. A competence which is connected to vigilance and awareness of the tendencies and complexities of AI-systems that put fundamental human rights at stake, including the right to privacy and data protection,

[221] UNESCO Recommendations on AI Ethics, 33.

[222] UNESCO, *K-12 AI curricula: a mapping of government-endorsed AI curricula* (n12).

[223] Ibid.

[224] Ibid. 10-11.

[225] E.g. UNESCO Recommendations on AI Ethics (n 6).

[226] Ibid. 22.

freedom of thought and expression, fairness and non-discrimination.[227] It represents a holistic approach mirrored by what the AI Act demands.

10.3 The EU AI Act

The EU AI Act[228] constitutes the first horizontal regulation of AI in history.[229] It embodies a risk-based classification system (prohibited, high-risk, limited and minimal risk) to attach proportionate obligations aimed at system providers and deployers.[230] Its objective is to promote the uptake of trustworthy and human-centric AI while protecting individuals safety, health and fundamental rights and wider Union values.[231] It is perceived as a key policy effort in emphasizing the need of AI-literacy, harnessing AI-related risks, and opportunities; particularly associated with systems classified as high-risk.[232]

Notably, the Act's approach goes beyond a product safety legislative framework on which it was modelled.[233] It is not simply designed to hold manufacturers and providers liable for their product's faults (here, AI-systems'), complaints and harm to health or safety *post-factum*. It is a

---

[227] See principles ibid. 20-23.

[228] 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 (COM/2021/206).

[229] Tambiama Madiega, 'Artificial intelligence act' (2021) European Parliament: European Parliamentary Research Service, 1.

[230] Yasmina Yakimova and Janne Ojamo, 'Artificial Intelligence Act: MEPs adopt landmark law' *European Parliament Press Release* (13 March 2024) <https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law>; Corrigendum (n 1), rec 26.

[231] Corrigendum (n 1), rec 1.

[232] Villar Onrubia Daniel and others, 'On the Futures of Technology in Education: Emerging Trends and Policy Implications' (2023) 38.

[233] Madiega (n 22), 12 endnote 11.

framework that also assigns responsibility for (a) implementing appropriate measures of redress to violations of a fundamental rights nature (atypical to product safety legislation),[234] and (b) gives rise to individual subject's rights (e.g. right to an explanation).[235] However, most importantly, for this paper, the Act is also designed to hold providers liable for (c) implementing and maintaining a comprehensive, *proactive* risk management system through precautionary measures such as human oversight.[236]

10.4 (Human) risk management measure

Establishing a Risk Management System is predominantly directed at the providers;[237] however, it is observed that certain consequent legal obligations are allocated across a variety of players (including deployers, the Board, and Office).[238] They bear joint responsibility for implementing or supporting the instructions and measures prepared by the providers.[239] The Act's distinct approach account's for the *complexity* of AI-systems supply chain, which often diffuse the chain of command over the particular deployment and use of AI-systems output.[240]

---

[234] Corrigendum (n 1), art 1.

[235] Ibid. art 86; akin to Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data ('General Data Protection Regulation', 'GDPR') Recital 71.4.

[236] Corrigendum (n 1), art 8, 9.

[237] Corrigendum (n 1), art 17(1)(g).

[238] For Supply Chain see e.g. Corrigendum (n 1) art 25; Deployers art 26; Office e.g. art 95, 62(3), 27(5); European Artificial Intelligence Board e.g. art 66.

[239] Ibid. art 17 (1).

[240] Committees European Parliament, 'REPORT on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules

Notably, following the original EP Mandate Recital 43[241] (now Recital 66),[242] the required HRAIS risk management systems are evaluated against the *relevance* and *quality* of measures and design of datasets, technical documentation, information provision, and human oversight etc. Arguably, these two *risk mitigation* features characterize what measures are required to satisfy human oversight obligations. The obligations associated with effectively safeguarding health, safety and fundamental rights of individuals from potential adverse effects of HRAIS.[243]

Human Oversight ('HO'), can therefore be interpreted as a Risk Management Measure,[244] within a wider context of a Risk Management System. Whereas 'AI-literacy' can be seen as a building block of the HO measure, that should be assessed for its *necessity* and *quality* in fulfilling specific HO obligations.

Accordingly, the paper associates (i) necessity and (ii) quality of the provision – with the role of 'AI-literacy' in fulfilling HO obligations (i.e. *necessity element*) and with what a desired level of 'AI-literacy' entails (i.e. *quality element*). This de-factorization is used to answer the question at hand. Namely, whether the 'AI-Literacy' mandate is sufficient to ensure that Human Overseers can fulfil their role as per Article 14. It is argued that the AI-Literacy mandate has several implications;

---

on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts' (2023) A9-0188/2023 p. 367.

[241] European Commission, ' 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 (Tabular Ammendments)' (21 January 2024) 2021/0106(COD).

[242] Corrigendum (n 1).

[243] Ibid. art 14(2).

[244] Ibid. art 14(3).

but, whether it is adequate to *enable* effective HO will depend on their necessity and quality.

10.5 Evolution of AI-literacy: Quality and necessity

The evolution of the AI-literacy mandate will be conducive to understand its final construction, its necessity and quality as it is juxtaposed with its preceding versions. 'AI-literacy' provisions were first officially exposed through Parliamentary Amendments, released on 22nd May 2023 in the Report on the Proposal.[245] Seeking to amend the initial Proposal of the Commission,[246] two co-rapporteurs, needed to accommodate "3,300 amendments tabled by different political groups",[247] before the joint deliberations with Committee on Internal Market and Consumer Protection ('IMCO') and the Committee on Civil Liberties, Justice and Home Affairs ('LIBE').[248]

Whilst the mandate's precise origin in the Act is unclear, it has received a lot of attention from different Parliamentary Committees before the report was officially released.[249] Most notably, from the Committees on (i) Industry, Research and Energy (ii) Culture and Education and (iii) Legal Affairs (contained in the May 2023 Report).[250] Based on publicly

[245] European Parliament (n 33), May Report A9-0188/2023.

[246] European Commission, Tabular Ammendments (n 24).

[247] Luca Bertuzzi and Molly Killeen, 'Tech Brief: AI Act amendments, Germany's data retention, standardisation politics' (*EurActiv,* 3 June 2022) <https://www.euractiv.com/section/digital/news/tech-brief-ai-act-amendments-germanys-data-retention-standardisation-politics/> accessed 3 March 2024

[248] Luca Bertuzzi, 'AI regulation filled with thousands of amendments in the European Parliament' (*EurActiv,* 7 June 2022) <https://www.euractiv.com/ section/digital/news/ai-regulation-filled-with-thousands-of-amendments-in-the-european-parliament/> accessed 3 March 2024

[249] European Parliament (n 33), May Report A9-0188/2023, 369-529.

[250] Ibid. 369-529.

available information, the downstream construction can be traced back to the influential opinions of the three, out of five Parliamentary Committees.[251] These, have provided an explicit opinion on how the 'AI-literacy' mandate should be understood, as a means of fulfilling HO sufficiently.

The first, Committee on Industry, Research and Energy, predominantly focused on the role of 'AI-literacy' as the means necessary for regulatory compliance and enforcement (*necessity element*).[252] However, they have not elaborated beyond the initial construction of the *quality* element, encompassing the "skills, knowledge and understanding regarding AI-systems".[253]

The second, representing the interests of Culture and Education sectors, have proposed several points associated with AI-literacy's quality and necessity. Particularly, in their proposed amendments for recitals 14b and art.4b, they have characterized AI-literacy to entail critical thinking skills and knowledge of basic notions of AI-systems, their functioning, applications (tools and technologies), benefits and risks (e.g. manipulative and harmful uses), their severity and how probable they are.[254] However, the Committee did not constrain the term to operators-only; they encompassed citizens (of all society sectors, ages and especially women), who should be allowed to make informed use and deployment of these systems, and their opportunities for public good.[255]

Importantly, they also vividly drew the *necessity* element of AI-literacy, as an unavoidable component allowing operators and deployers to comply (or to justify non-compliance) with articles on appropriate HO,[256]

---

[251] Ibid.

[252] Ibid. 390.

[253] Ibid.

[254] European Parliament (n 33), May Report A9-0188/2023, 423, 439, 440.

[255] Ibid.

[256] European Commission, Tabular Ammendments (n 24), art 29 (1a).

information transparency,[257] and biometric identification.[258] Distinctly, they proposed to include the "provision of a sufficient level of AI-literacy" as a part of providers' compulsory Risk Management System.[259] Where, 'sufficient' level of literacy depends on whether it equips operators to "fully comply with and benefit from trustworthy AI", and specifically whether it enables them to comply with the Regulation.[260]

Finally, the Committee on Legal Affairs, also characterized the *quality element* of AI-literacy. It proposed that *sufficient* literacy encompasses not only the knowledge, skills and understanding of AI-systems (their tools and technologies), but equally includes the awareness of respective obligations and rights arising from the Regulation.[261] S*ufficient* literacy to enable providers, deployers – and – affected persons, to make informed choices about AI-systems deployment and having the awareness of its risks, opportunities and resultant harms to facilitate democratic control.[262] It contended that AI-literacy should also equip these three groups with the skills and notions to ensure Regulation enforcement and compliance.[263] The Committee reiterated that it is a crucial measure to accomplish providers' and deployers' responsibilities triggered by Art.13 (on information transparency),[264] and chiefly, the HO function (*i.e. necessity element*).[265]

In hindsight, the common themes, include qualifying the sufficiency of AI-literacy by considering whether or not it enables or ensures

[257] Ibid. art 13.
[258] Ibid. art14(5); c.f. Corrigendum (n 1) art14(5).
[259] European Parliament (n 33), May Report A9-0188/2023, 442.
[260] Ibid. 423, 439, 440.
[261] Ibid. 460-461, 477.
[262] Ibid. 479-480.
[263] Ibid. 460-461.
[264] European Commission, Tabular Ammendments (n 24), art 13.
[265] European Parliament (n 33), May Report A9-0188/2023, 494.

compliance; and encompasses skills, knowledge and understanding of these systems by all citizen groups and supply chain actors. The necessity element was common across the Legal Affairs and Culture and Education Committees emphasizing AI-literacy's role in compliance with various provider and deployer obligations – especially related to HO and transparency.

Even though the three opinions seem to diverge, many elements were accommodated for in the May Report.[266] Notably, the co-rapporteurs characterized a *sufficient* level of AI-literacy to include skilling and reskilling of all three groups (as identified by the Legal Affairs Committee), on basic notions of AI-systems functioning, applications, uses, their benefits and risks.[267] Alongside a wider commitment of promoting AI-literacy of all sectors, and people of all ages.[268] It also reflected the *necessity* element as advocated by all three Committees. Namely, that AI-literacy is necessary to enable deployers and providers to *ensure* regulatory compliance and enforcement.[269] It considered this in the context of the obligations across the supply chain (i.e. the Board,[270] AI Office,[271] and Member States).[272]

This May 2023 version has indeed proposed a much more comprehensive, detailed understanding of 'AI-literacy' than what now constitutes the final version of the Proposal, which has recently undergone a final lawyer-linguist check.[273] Chiefly, both overt and covert negotiations have resulted in numerous compromises on what now constitutes the final

[266] European Parliament (n 33), May Report A9-0188/2023.
[267] Ibid.
[268] Ibid. Rec 9(b).
[269] Ibid.
[270] Ibid. art 58.
[271] Ibid. art 69(2).
[272] Ibid. art 4(b).
[273] Corrigendum (n 1).

text. The publicly available amendments, and deliberations were exposed through a tabular juxtaposition of the Original Commission Proposal, alongside the EP Mandate, the Council Mandate, which resulted in the final column containing the agreed draft.[274] These final-text provisions contained in the "corrigendum cor01",[275] form the latest available document awaiting the Council's formal endorsement.[276]

10.6 Final version 'AI-literacy' mandate

The final version of the 'AI-literacy' mandate is substantially different from its proposed construction. It became much more narrow, but arguably retains important detail. Crucially, the AI-literacy provisions still stand as an important building block supporting the HO framework, and still *indirectly* constitutes a key ingredient of regulatory compliance. Nonetheless considering the changes – the current scope of the mandate, and how the final-text builds on the notions of (i) necessity and (ii) quality must now be clarified.

AI-literacy as a legal requirement was redrafted to only target the Providers and Deployers directly involved in the use and operation of AI-systems under Article 4, rather than all three groups identified in the EP Mandate (i.e. including affected persons).[277] Instead, the language of "promoting" public awareness was used when imposing general AI-literacy-related obligations on the Board, and the AI Office.[278] However, the AI-literacy "promotion" obligation of the Office, was also

[274] See European Commission, Tabular Ammendments (n 24) Final Draft (21 January 2024) 17.11.

[275] Corrigendum (n 1).

[276] Yakimova and Ojamo, (n 23).

[277] Corrigendum (n 1) art 4; c.f. European Parliament (n 33), May Report A9-0188/2023 art 4b.

[278] Corrigendum (n 1) art 66 (1)(f), art 95(2)(c).

constrained to those involved in the development, operation and use of AI-systems.[279] Similarly, no target audience was explicitly identified in the Board's obligations.[280] This arguably communicates, a notable change in the priorities of the AI Act's literacy mandate.

The *necessity* element of AI-literacy has also substantially shifted. It is no longer *explicitly* mentioned as a binding measure designed to ensure or enable regulatory compliance of providers and deployers with HO and transparency obligations throughout.[281] Nonetheless, this paper purports that due to the background evolution and strong emphasis of the negotiating parties, it is likely to continue to perform this function. An indicative example includes a trace of the old construction which was retained under the non-binding Recital 91. It suggests that 'AI-literacy' is still *necessary* to *ensure* that individuals (assigned with HO, or the implementation of providers' instructions for AI-systems' use) can *properly* fulfill their tasks.[282]

The *necessity* element (i.e. the role of AI-literacy in fulfilling regulatory compliance), can also be inferred from the consequent obligations of providers to ensure that HRAIS satisfy Section 2 requirements.[283] This includes Art 8 and 9,[284] which mandates providers to ensure compliance through an appropriate risk management system, which necessitates HO measures,[285] which in turn can be linked to AI-literacy under Art.4. A chain of responsibilities, indirectly communicating that AI-

[279] Ibid. art 95(2)(c).

[280] Ibid. art 66(1)(f).

[281] C.f. European Parliament (n 33), May Report A9-0188/2023 e.g. art 4b (4), and art 13 (3a).

[282] Corrigendum (n 1) rec 91.

[283] Ibid. art 16(1)(a).

[284] Ibid. art 8(1); art 9.

[285] Ibid. art14.

literacy is *still necessary* for regulatory compliance with, at least, the HO obligations.

The *quality* element (i.e. what AI-literacy entails) continues to relate to a *sufficient* level of AI-literacy of those involved in AI-systems use and operation.[286] The level, is commensurate to their training, education, experience and technical knowledge and the systems' context.[287] It should entail, alongside general awareness, an understanding of the benefits, risks and safeguards, obligations and rights related to the system's use.[288] Whilst these aspects stem from the binding articles, the non-binding recitals add, that AI-literacy should also correspond to the knowledge and skills allowing *all* parties (i.e. deployers, providers *and* affected individuals) to make informed AI-systems deployment, and to be aware of the opportunities, risks and possible harms of AI-systems.[289]

Recital 20, also contains the lost EP Mandate's recommendations, to measure the *quality* and *sufficiency* of AI-literacy, by asking whether it offers the "relevant actor in the AI value chain (…) the insights" to ensure regulatory compliance and enforcement.[290] This recital also connects literacy with the means to achieving the "greatest benefits from AI-systems" to fulfil the wider objectives of the Regulation; namely, safeguarding health, safety and fundamental rights of persons.[291] Chiefly, it clarifies that literacy, is qualified by "necessary notions" – which can include the understanding of measures to be applied during AI-systems use, the technical elements' correct application (when developed), and the knowledge of appropriate ways to interpret AI-systems output.[292]

[286] Corrigendum (n 1) art 4.
[287] Ibid. art 4.
[288] Ibid. art 66(1)(f).
[289] Ibid. rec 56.
[290] Ibid. rec 20.
[291] Ibid. rec 20.
[292] Ibid. rec 20.

It is, notably, a result of a compromise reached during the trialogue, and the future weight assigned to this recital may ultimately determine what the legal obligation of AI-literacy of human overseers entails or not.

Given the premises of the binding (i) *quality* and (ii) *necessity* of AI-literacy provisions in the context of the Act– the paper may now specifically ask whether it is *sufficient* to ensure that Human Overseers can fulfill their Art.14 obligations. This will require a closer look at how HO obligations should be met, and accordingly what *literacy* is called for.

10.7 Meeting human oversight obligations

Human oversight and agency is interwoven with developing AI-systems that are used as a tool that serves and respects people, their dignity and autonomy through appropriate control and monitoring.[293] Article 14 obligations of those entrusted with fulfilling the oversight role involves onerous responsibilities to effectively oversee and facilitate the correct implementation of risk management measures, to eliminate or minimize the likelihood of harm.[294] Oversight is ensured through measures that are either identified and built by providers into the HRAIS, or identified by the provider and implemented by the deployer.[295]

Ultimately, proportionate measures should enable natural persons to: (a) understand relevant limitations and capacities of HRAIS properly, to monitor its operation; (b) remain vigilant of the tendency to over-rely or rely on system output (i.e. automation bias) (c) interpret HRAIS output correctly (d) decide when to disregard, override or reverse its output and (e) when to interrupt (safely halt) the system.[296] As recitals explain, it

[293] Corrigendum (n 1) rec 27.

[294] Ibid. art 14(2).

[295] Ibid. art 14(3).

[296] Ibid. art 14(4). (main obligations)

constructs a position, which necessitates training and authority and *competence*,[297] including a sufficient level of AI-literacy.[298] It is a duty that depends on the wider organizational and technical measures taken by deployers to ensure that HRAIS can be used, and HO implemented, in accordance with providers' instructions.[299] Hence, these obligations and their implications shape the role of Human Overseers as per Art.14, and signify what 'literacy' is called for.

What literacy these obligations ask for, may be proven by looking at the contrary; namely, situations where a lack of literacy is equivalent with failing regulatory obligations. These are subsequently discussed.

Firstly, AI Act's main aim is to encourage the uptake of responsible and ethical AI. A lack of skills may constitute a major barrier to adopt these technologies,[300] as reiterated by European Commissions Joint Research Centre.[301] Hence, providing appropriate support and training in AI-literacy, is likely to be pivotal to fulfil the objectives of the framework. To enable those assigned to fulfil and take on these duties in the first place.

Secondly, HO provision's purpose is to minimize, or prevent risks to health, safety and fundamental freedoms.[302] However, the lack of literacy, as EP's Science Panel highlight, can become the operative cause for AI misuse by domain-experts, such as clinicians.[303] They report that the

---

[297] Ibid. rec 17.

[298] Ibid. rec 91.

[299] Ibid. art 13 (3), rec 91.

[300] OECD (n 13) *Employment Outlook 2023*, 168.

[301] Lorenzino Vaccari and others, 'AI watch, beyond pilots : sustainable implementation of AI in public services' (2021) 30.

[302] Corrigendum (n 1) art 14(2).

[303] Science Panel for the Future of and Parliament Technology for the European, 'Artificial intelligence in healthcare: Applications, risks, and ethical and societal

lack of training and explanation of the system, can contribute to the inappropriate use of widely-accessible AI, and consequently a lack of appropriate oversight.[304] Which, in turn, can undoubtedly exponentiate the likelihood of harm, especially in critical sectors such as healthcare.

Thirdly, AI-literacy, should enable Human Overseers to interpret the HRAIS output.[305] The lack of appropriate literacy for HO, may constitute barriers to abide by demanding obligations, such as explaining the AI-systems output correctly.[306] While not all AI-systems, may require the use of XAI techniques to devise an appropriate explanation, or interpret the systems output correctly – some are likely, particularly in view of their complex nature, and lack of system interpretability (e.g. XAI for Neural Networks).[307]

These are just three illustrative examples, where the lack of AI-literacy, may de facto, hinder regulatory compliance. Meanwhile, the paper acknowledges that knowledge and understanding fostered by upskilling or reskilling workers in AI-literacy only,[308] may not be enough to fulfil the aforementioned HO obligations. This is because they may necessitate other, distinct forms of competence or technical literacy, such as XAI-literacy, or HCI competences to be able to engage critically and interact appropriately with machines and technologies where AI-systems are embedded. This may flow from human-machine interface tools,[309] or

---

impacts | Panel for the Future of Science and Technology (STOA) | European Parliament' (2022) 6.

[304] Ibid.

[305] Corrigendum (n 1) art 14(4)(c).

[306] Ibid. art 14(4)(c); *akin to* GDPR (n 28) art 22, rec 71.

[307] Information Commissioner's Office and The Alan Turing Institute, *'Explaining decisions made with AI' Project ExplAIn* (2022), 21-35 (see explanations); 69-70 (see black-box systems); 74-75 (see techniques).

[308] Cristiano Codagnone and others, 'Ethics in the digital workplace' (2022) 25.

[309] Corrigendum (n 1) art 14(1).

interpretability mechanisms,[310] that may be particularly complex to operate or oversee.

Arguably, what is necessary to sufficiently fulfil HO obligations, is a *holistic* approach to prepare those assigned with HO to carry out their role. What UNESCO outlined as not simply AI-literacy – but AI *competence* (See above 10.2). Which should account for building skills, knowledge, understanding and value orientation of those fulfilling the Art.14 function. While AI-literacy is not compromised in importance; s*ufficient competence*, should be fostered through organizational and technical infrastructures. These could be achieved through the support and training put in place by the providers and deployers, to help HO acquire necessary, AI-system specific skills.

Accordingly, the requisite level of *literacy* in AI-literacy, is observed to go beyond AI-related knowledge. It could be compared to the field of data science, which requires interdisciplinary insights from statistics, mathematics, programming and, importantly, *domain-knowledge* to yield appropriate insight from data.[311] Similarly, as emphasized by earlier policy documents (above 10.2), training in wider civic responsibility and ethics may also constitute a key element of *AI competency* when deploying and overseeing complex technologies responsibly.[312] It follows, that AI-literacy alone may not be sufficient to achieve the complex legal aspirations underpinning Article 14. It must be understood in the context of AI-systems deployment, the intended application domain, the existing infrastructure and responsibilities of respective supply chain actors. Moreover, AI-literacy, as a 'pillar' of AI competence, should be

[310] Ibid. art 14(4)(c).

[311] Whitehouse.gov, *NSTC: Building Computational Literacy Through STEM Education: A Guide for Federal Agencies and Stakeholders* 26.

[312] Ibid.

understood in the shadow of its policy formulation – to give full effect to what it was designed to encompass.

10.8 Conclusion

As natural persons are entrusted with Human Oversight, they undertake the role of guardians of AI-systems safe, intended deployment and the mechanisms protecting rights at risk. AI-literacy, may be vital in the ethical and responsible AI uptake, by not only understanding AI-related concepts, but being able to assess its trustworthiness, and the reliability of the tools' infrastructure.[313] However, AI-literacy need not be automatically equated with the ability to develop AI models.[314] Notably, the human-oversight mandate does not require a formal qualification or certification to fulfil these obligations, but *sufficient competence*, which is likely to rely on 'AI-literacy', training and support, rather than 'AI expertise' *per se*. Chiefly, with the proliferation of AI tools among all sectors, training non-experts becomes as important as upskilling AI experts – as we learn to live with it.[315]

Accordingly, to address the question 'whether AI-literacy is sufficient to ensure that Human Overseers can fulfil their role as per Article 14' – this paper argued that while AI-literacy is undoubtedly a necessary element in the equation, it strongly depends on two notions of (i) necessity and (ii) quality as construed by the law. While the mandate's evolution exposes what the two characteristics entail – how this 'pillar' holds, will be relative to its implementation and the AI-systems' context of use.

---

[313] Knaw, 'Successful and timely uptake of Artificial Intelligence in science in the EU' Expert Report (European Commission Scientific Advice Mechanism 2024), 22.

[314] IZA Institute of Labor Economics, *Artificial Intelligence Capital and Employment Prospects*, 2024) 2.

[315] IZA Institute of Labor Economics (n 107).

In 2023, the OECD reported the lack of constructive policies which propose sufficient, *concrete* actions to develop such AI skills and relevant literacy.[316] What remains to see, is whether a change will follow the AI Act. It may well be that the development of precise guidance, certification mechanisms and standards regarding AI-literacy and *competence*,[317] will become the cornerstone of achieving the law's *propositum* behind human oversight obligations – regardless of how 'AI-literacy' itself is tentatively discussed.

Table of authorities

Primary law instruments

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 (COM/2021/206)

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data ('General Data Protection Regulation', 'GDPR').

Policy documents

European Commission, ' 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 (Tabular Ammendments)' (21 January 2024) 2021/0106(COD)

[316] OECD (n 13), *OECD Employment Outlook 2023*, 170-172.

[317] Commission European and others, *Analysis of the preliminary AI standardisation work plan in support of the AI Act* (Publications Office of the European Union 2023) 11-13.

European Parliament and Committees, 'CORRIGENDUM to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts ('Corrigendum')' (19 April 2024) cor01

European Parliament and Committees, 'REPORT on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts' (2023) A9-0188/2023

Policy reports and guidance

Canada. Innovation S, Economic Development Canada ib and Government of C, 'Learning together for responsible artificial intelligence : report of the public awareness working group.: Iu173-38/2022E-PDF Publications - Canada.ca' (2022)

Codagnone C and others, 'Ethics in the digital workplace' (2022)

Council of the European Union, 'Council Recommendation on the key enabling factors for successful digital education and training' (2023)

Daniel VO and others, 'On the Futures of Technology in Education: Emerging Trends and Policy Implications' (2023)

European Commission and others, *Analysis of the preliminary AI standardisation work plan in support of the AI Act* (Publications Office of the European Union 2023)

European Education Culture Executive Agency, *'Informatics education at school in Europe*: Eurydice report' (Publications Office of the European Union 2022)

European Parliament Plenary, 'European Parliament resolution of 21 November 2023 with recommendations to the Commission on an EU framework for the social and professional situation of artists and

workers in the cultural and creative sectors (2023/2051(INL))' (2023)

Information Commissioner's Office and The Alan Turing Institute, *'Explaining decisions made with AI' Project ExplAIn* (2022)

Knaw, 'Successful and timely uptake of Artificial Intelligence in science in the EU' Expert Report (European Commission Scientific Advice Mechanism 2024)

Madiega T, 'Artificial intelligence act' (2021) European Parliament: European Parliamentary Research Service

Panel for the Future of S and Technology for the European P, 'Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts | Panel for the Future of Science and Technology (STOA) | European Parliament' (2022)

Science Technology, Committee U. K. Parliament Select Committee Publications, 'GAI0084 - Governance of artificial intelligence (AI)' (18 July 2022)

UNESCO Recommendation on the Ethics of Artificial Intelligence ('UNESCO Recommendations on AI Ethics') (Paris, 2022), SHS/BIO/PI/2021/1

Whitehouse.gov, 'NSTC: Building Computational Literacy Through STEM Education: A Guide for Federal Agencies and Stakeholders' (2023)

Vaccari L and others, 'AI watch, beyond pilots: sustainable implementation of AI in public services' Joint Research Centre (Publications Office of the European Union 2021)

Bibliography

Articles

Long D and Magerko B, 'What is AI literacy? Competencies and design considerations' (2020) Proceedings of the 2020 CHI conference on human factors in computing systems Ng DTK and others, 'Conceptualizing AI literacy: An exploratory review' (2021) 2 Computers and Education: Artificial Intelligence 100041

Reports

I. Z. A. Institute of Labor Economics, *Artificial Intelligence Capital and Employment Prospects*, (2024)

Eric S, Brian A and OECD, *Collective action for responsible AI in health*, (2024)

European, Schoolnet, *Data4Learning: ChatGPT and the role of AI in assessment* (2023)

OECD, *OECD Employment Outlook 2023* (2023)

UNESCO, *K-12 AI curricula: a mapping of government-endorsed AI curricula* (2022)

Websites

Bertuzzi L, 'AI regulation filled with thousands of amendments in the European Parliament' (*EurActiv,* 7 June 2022) <https://www.euractiv.com/section/digital/news/ai-regulation-filled-with-thousands-of-amendments-in-the-european-parliament/> accessed 3 March 2024

Bertuzzi L and Killeen M, 'Tech Brief: AI Act amendments, Germany's data retention, standardisation politics' (*EurActiv,* 3 June 2022) <https://www.euractiv.com/section/digital/news/tech-brief-ai-act-amendments-germanys-data-retention-standardisation-politics/> accessed 3 March 2024

Yakimova Y and Ojamo J, 'Artificial Intelligence Act: MEPs adopt landmark law' *European Parliament Press Release* (13 March 2024) <https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

PART II

AI GOVERNANCE:
COUNTRY-SPECIFIC PERSPECTIVES

11

CHINA'S EFFORTS AND APPROACHES TO AI GOVERNANCE

Zheng LIANG and Chang LIU, China

11.1 Characteristics of AI development in China

11.1.1 Strong research base

China[318]has a long history of AI research and has become a world leader in AI publications. According to Stanford HAI's AI Index Report 2024, the academic sector contributed most AI publications (81.1%) in 2022, maintaining its position as the leading global source of AI research over the past decade across all regions.

China contributed the most in this sector, representing 81.75% (Stanford Institute for Human-Centred Artificial Intelligence 35). This strong

---

[318] LIANG Zheng serves as Professor of the School of Public Policy and Management, Tsinghua University, Beijing, and Vice Dean of its Institute for AI International Governance (I-AIIG). He is also Research Fellow and Deputy Director of the China Institute for Science and Technology Policy at Tsinghua University (CISTP).

LIU Chang, PhD, graduated in Singapore with a Master of Arts. She is the coordinator of Foreign Affairs of the Institute for AI International Governance of Tsinghua University. She once worked at the Chinese Institute of Electronics, Cross Ratio (Singapore), and the Secretariat of S&T-oriented International Organizations.

research background fuels innovation and contributes to the development of cutting-edge AI technologies and applications.

11.1.2 Broad application

The World Intellectual Property Organization's *Patent Landscape Report on Generative AI* reveals that China produced more than 38,000 Generative AI inventions between 2014-2023, six times more than the second-place U.S. (World Intellectual Property Organization 40). Generative AI is already spreading across industries, including the life sciences, manufacturing, finance, and transportation in China. Moreover, Stanford HAI's *AI Index Report 2024* shows that the total AI private investment in China was 103.65 U.S. dollars between 2013-2023, ranking second globally (Stanford Institute for Human-Centered Artificial Intelligence 248).

11.1.3 Close linkage between development and governance

While promoting AI research, development, and investment, China frequently explores options for AI technology regulation. From 2017 to the present, China has issued thirteen documents, including laws, regulations, guidelines, and initiatives to guide and regulate AI's responsible and ethical development.

11.1.4 Rapid growth in Large-Language Models (LLMs)

The year 2020 marked the beginning of China's accelerated growth in the field of LLMs. According to the *China Large-scale Language Model Map Research Report* released by the New Generation Artificial Intelligence Development Research Center of the Ministry of Science and Technology in May 2023, Chinese organizations launched 79 LLMs between 2020-2023 (Moraitis). Globally, China and the United States lead in LLM development, accounting for more than 80% of the total (Moraitis).

11.1.5 Full engagement in AI global governance

China has actively participated in major global initiatives and dialogues of AI governance, with the *Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence* and the *Global AI Governance Initiative* launched in recent years. Each of these demonstrates China's advocation of the principles of wide participation and consensus-based decision-making in global AI governance and its efforts to promote broad international consensus based on full respect for differences in policies and practices of all countries.

11.2 China's journey in building an AI governance system

11.2.1 Eight governance principles

Under the philosophy of "People-Centered and AI for Good," China has been developing an AI governance system since 2017, with the *Cybersecurity Law of the People's Republic of China* taking effect in June 2017. The law safeguards cyberspace sovereignty and national security, as well as social and public interests, and promotes the healthy development of the information of the economy and society (Standing Committee, "Cybersecurity Law"). Moreover, the *New Generation Artificial Intelligence Development Plan* issued in July 2017 stresses that the dual technical and social attributes of AI must be carefully managed to ensure that AI is trustable and reliable (State Council).

In June 2019, the National Governance Committee for the New Generation Artificial Intelligence, a body established in February 2019 to address the legal and ethical issues surrounding AI use, issued a document outlining eight governance principles for the development of AI. Titled "Governance Principles for the New Generation Artificial Intelligence" with a subtitle of "Developing Responsible Artificial Intelligence," this

document delineates the following *eight points as "Governance Principles"*:

1. Harmony and human-friendly: The goal of AI development should be to promote the well-being of humankind. It should conform to human values and ethical principles, promote human-machine harmony, and serve the progress of human civilization. The development of AI should be based on the premise of ensuring social security and respecting human rights, and should prevent misuse, abuse, and evil use of AI technology by all means.

2. Fairness and justice: The development of AI should promote fairness and justice, protect the rights and interests of all stakeholders, and promote equal opportunities. Through technology advancement and management improvement, prejudices and discriminations should be eliminated as much as possible in the process of data acquisition, algorithm design, technology development, and product development and application.

3. Inclusion and sharing: AI should promote green development to meet the requirements of environmental friendliness and resource conservation; AI should promote coordinated development by promoting the transformation and upgrading of all industries, and by narrowing regional disparities; AI should promote inclusive development through better education and training, support to the vulnerable groups to adapt, and efforts to eliminate digital divide; AI should promote shared development by avoiding data and platform monopolies, and promoting open and fair competition.

4. Respect for privacy: AI development should respect and protect the privacy of individuals and fully protect an individual's rights to know and to choose. Boundaries and rules should be established for the collection, storage, processing, and use of personal information. Personal privacy authorization and revocation mechanisms should be established and

updated. Stealing, juggling, leaking, and other forms of illegal collection and use of personal information should be strictly prohibited.

5. Safety and controllability: The transparency, interpretability, reliability, and controllability of AI systems should be improved continuously to make the systems more traceable, trustworthy, and easier to audit and monitor. AI safety at different levels of the systems should be ensured, AI robustness and anti-interference performance should be improved, and AI safety assessment and control capacities should be developed.

6. Shared responsibility: AI developers, users, and other related stakeholders should have a high sense of social responsibility and self-discipline, and should strictly abide by laws, regulations, ethical principles, technical standards, and social norms. AI accountability mechanisms should be established to clarify the responsibilities of researchers, developers, users, and relevant parties. Users of AI products and services and other stakeholders should be informed of the potential risks and impacts in advance. Using AI for illegal activities should be strictly prohibited.

7. Open and collaboration: Cross-disciplinary and cross-boundary exchanges and cooperation should be encouraged in the development of AI. Coordinated interactions should be fostered among international organizations, government agencies, research institutions, educational institutions, industries, social organizations, and the general public in the development and governance of AI. With full respect for the principles and practices of AI development in various countries, international dialogues and cooperation should be encouraged to promote the formation of an international AI governance framework with broad consensus.

8. Agile governance: The governance of AI should respect the underlying principles of AI development. In promoting the innovative and healthy development of AI, high vigilance should be maintained in order

to detect and resolve possible problems in a timely manner. The governance of AI should be adaptive and inclusive, constantly upgrading the intelligence level of the technologies, optimizing management mechanisms, and engaging with muti-stakeholders to improve the governance institutions. The governance principles should be promoted throughout the entire lifecycle of AI products and services. Continuous research and foresight for the potential risks of higher level of AI in the future are required to ensure that AI will always be beneficial for human society.

Since then, China has accelerated the exploration of building an AI governance system with a series of document releases targeting specific types of algorithms and AI capabilities to guide AI governance development, including codes of conduct, ethical norms, technical standards, laws and regulations, and international initiatives:

11.2.1 Main documents for AI governance in China

Table 1. Main documents for AI governance in China

Document	**Issued**	**Brief introduction**
Guidelines for the Construction of a National New Generation Artificial Intelligence Standards System	27 July 2020	This document calls for the formulation of specific standards across the full spectrum of basic and applied AI technologies. The standards focus on the fields of natural language processing and speech-related applications of AI (Standardization Administration of China et al., "Guidelines").
Data Security Law of the People's Republic of China	10 June 2021	This law aims to standardize data handling activities, ensure data security, promote data development and use, protect the lawful rights and interests of individuals and

		organizations, and safeguard national sovereignty, security, and development interests. The law entered into effect on 1 Sept. 2021 (Standing Committee, "Data Security Law").
Ethical Norms for New Generation Artificial Intelligence	25 Sept. 2021	This document outlines ethical norms for the use of AI in China. The norms address areas such as the use and protection of personal information, human control over and responsibility for AI, and the avoidance of AI-related monopolies (National Governance Committee, "Ethical Norms").
Personal Information Protection Law of the People's Republic of China	20 Aug. 2021	This law is designed to protect the rights and interests of personal information, regulate personal information processing activities, and promote the rational use of personal information. The law took effect on 1 Nov. 2021 (Standing Committee, "Personal Information Protection Law").
Provisions on Administration of Algorithm-generated Recommendations for Internet Information Services	31 Dec. 2021	These provisions aim to standardize Internet information service algorithmic recommendation activities, safeguard national security and social and public interest, and protect the lawful rights and interests of

		citizens, legal persons, and other organizations. They entered into force on 1 Mar. 2022 (Cyberspace Administration of China et al., "Provisions").
Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence	16 Nov. 2022	This paper highlights China's vision, practices, and views on ethical governance of technology from the perspective of global cooperation and coordination. It calls for global consensus with mutual respect and actions for the good of humanity (Ministry of Foreign Affairs, "Position Paper").
Interim Measures for the Management of Generative Artificial Intelligence Services	13 July 2023	This document outlines a set of measures introduced by China to regulate public-facing generative artificial intelligence within the country. These measures took effect on 15 Aug. 2023 (Cyberspace Administration of China, "Interim Measures").
Global AI Governance Initiative	18 Oct. 2023	This initiative centers on a people-centered approach, striving to elevate the well-being of humanity through AI development. It aims to address imbalances in AI capabilities among nations and calls on all countries to enhance information exchange and technological cooperation on the governance of AI

		(Ministry of Foreign Affairs, "Global AI Governance Initiative").
"AI Plus" Initiative	5 March 2024	This initiative aims to strengthen the connection between AI and various industries. This involves boosting research in practical AI applications, encouraging businesses to adopt AI with incentives, training skilled AI professionals, and building robust data infrastructure to support AI growth. By bridging the gap between AI and industries, this initiative will promote the innovative development of the digital economy.
Shanghai Declaration on Global AI Governance	4 July 2024	This declaration lays out China's position in five areas: (1) Promoting AI development; (2) Maintaining AI safety; (3) Developing the AI governance system; (4) Strengthening public participation and improving literacy; (5) Improving quality of life and increasing social well-being (Ministry of Foreign Affairs, "Global AI Governance Initiative").

China's AI governance system, guided by the philosophy of "People-Centered and AI for Good," is founded on eight governance principles and rooted in laws, technical regulations, ethical codes, and international standards. This system engages diverse stakeholders, employs

multidimensional governance approaches, and promotes agile collaboration. It reflects China's dedication to responsible AI development and its commitment to ensuring that technological advancements serve the interests of both individuals and society as a whole.

11.3 Stages of AI governance in China

Reviewing China's efforts in promoting AI governance, three distinct stages have come into focus, each representing a different *balance of interests and approaches to the issue*:

1. **Responsive regulations (2017-2020):** During this period, policies and measures focused on establishing a top-level governance structure. The emphasis was on providing incentives and encouragement for AI development and applications, promoting industry investment in the technology, and capitalizing on market opportunities. Policies were adjusted based on market feedback.
2. **Concentrated regulations (2020-2022):** This stage saw policies and measures that not only focused on basic legislation and targeted regulations but also intensified efforts to formulate industry standards. These standards aimed to cover algorithmic technologies and machine learning applications, protect public interests against risks in AI applications, and curb the market power of industrial giants.
3. **Agile regulation (2022-present):** In this current phase, policies and measures focus on maintaining and developing a regular governance framework, with coordination among different governance agencies and governance levels. A key aspect of this stage is the tailored approach Chinese regulators have adopted towards AI technologies. Each regulation currently in force has a more targeted scope compared to broader laws governing companies and technology providers.

11.4 Lessons learned from AI governance in China

China's AI governance system attempts to address multiple interconnected yet distinct issues originating from the technology, and this merits close attention for lessons learned:

1. **Science, technology, engineering, and math (STEM) capacity is crucial for AI development**: In recent years, China has gone to great lengths to emphasize the importance of STEM. STEM education fosters innovation, and its skills are essential for understanding the underlying mathematical and statistical concepts that underpin Machine Learning and AI algorithms.
2. **Agile governance is necessary**: The inherent uncertainty and rapid pace of AI technological development often outstrip the speed of policy formation and implementation. Consequently, the governance system must be prepared to respond quickly and adapt as needed. Agile governance enables the system to adapt to change and thrive in complex environments.
3. **Balance between different policy objectives is needed**: Striking a balance between promoting innovation and managing application risks is essential. It requires careful consideration of different stakeholders' perspectives and the implementation of various kinds of measures.
4. **Mutual adaptation between business and government is required**: Close collaboration and communication between the business sector and government are vital for identifying and addressing emerging risks that both innovators and regulators may overlook. Mutual understanding and self-adjustments between these sectors are needed for the integration of AI governance.
5. **Policy tools must be mixed**: AI governance should be clear about the general principles and directions and use mixed policy tools - guidelines, regulations, and laws. This approach aims not only to

manage risks but also to inspire and encourage industry innovation. Policymakers should be careful when using strong regulations and coordinate among governing agencies to avoid unintended compounding effects.

6. **Participation in global governance is encouraged**: AI governance is a global challenge that affects all of humanity, necessitating the establishment of global mechanisms to regulate high-risk scenarios. Global collaboration in AI research and governance is highly encouraged, and the building of inclusive global platforms is needed to promote communication and exchange.

References

Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security of the People's Republic of China, and State Administration for Market Regulation, "Provisions on Administration of Algorithm-generated Recommendations for Internet Information Services." *Cyberspace Administration of China,* https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm. (Accessed 17 July 2024).

Cyberspace Administration of China. "Interim Measures for the Management of Generative Artificial Intelligence Services." *Cyberspace Administration of China*, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. (Accessed 17 July 2024).

Ministry of Foreign Affairs. "Global AI Governance Initiative." *Ministry of Foreign Affairs of the People's Republic of China*, https://www.fmprc.gov.cn/eng/wjdt_665385/2649_665393/202310/t20231020_11164834.html. (Accessed 17 July 2024).

Ministry of Foreign Affairs. "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial

Intelligence." *Ministry of Foreign Affairs of the People's Republic of China*, https://www.fmprc.gov.cn/eng/wjb_663304/zzjg_663340/jks_665232/kjlc_665236/AI/202211/t20221117_10976730.html. (Accessed 17 July 2024.

Ministry of Foreign Affairs. "Shanghai Declaration on Global AI Governance." *Ministry of Foreign Affairs of the People's Republic of China*, https://www.fmprc.gov.cn/eng/wjdt_665385/2649_665393/202407/t20240704_11448349.html. (Accessed 17 July 2024)

Moraitis, Vangelis. "China's Rising Dominance in Large-Language Models (LLM)." *The AI Track*, https://theaitrack.com/hina-dominance-large-language-models/. (Accessed 17 July 2024)

National Governance Committee for the New Generation Artificial Intelligence. "Ethical Norms for New Generation Artificial Intelligence." *Ministry of Science and Technology of the People's Republic of China*, https://www.most.gov.cn/kjbgz/202109/t20210926 _177063.html. (Accessed 17 July 2024)

Standardization Administration of China, Cyberspace Administration of China, National Development and Reform Commission, Ministry of Science and Technology, and Ministry of Industry and Information Technology. "Guidelines for the Construction of a National New Generation Artificial Intelligence Standards System." *Gov. cn*, https://www.gov.cn/zhengce/zhengceku/2020-08/09/5533454/files/bf4f158874434ad096636ba297e3fab3.pdf. (Accessed 17 Jul. 2024)

Standing Committee of the National People's Congress of the People's Republic of China. *Cybersecurity Law of the People's Republic of China*. Nov. 2016.

Standing Committee of the National People's Congress of the People's Republic of China. *Data Security Law of the People's Republic of China*. June 2021.

Standing Committee of the National People's Congress of the People's Republic of China. *Personal Information Protection Law of the People's Republic of China*. Aug. 2021

Stanford Institute for Human-Centered Artificial Intelligence. *Artificial Intelligence Index Report 2024.* Stanford Institute for Human-Centered Artificial Intelligence Publishing, 2024. p. 35.

Stanford Institute for Human-Centered Artificial Intelligence. *Artificial Intelligence Index Report 2024.* Stanford Institute for Human-Centered Artificial Intelligence Publishing, 2024. p. 248.

State Council of the People's Republic of China. "New Generation Artificial Intelligence Development Plan." *The State Council of the People's Republic of China*, https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. (Accessed 17 July 2024)

World Intellectual Property Organization. *Patent Landscape Report on Generative AI*. World Intellectual Property Organization Publishing, 2024. p. 40.

"Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence." *China Daily*, https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html. (Accessed 15 July 2024).

Xue Lan attends the UNU Macau AI Conference 2024." *Institute for AI International Governance of Tsinghua University*, http://aiig.tsinghua.edu.cn/en/info/1025/1341.htm. (Accessed 18 July 2024)

12

OVERVIEW OF COUNTRY-SPECIFIC REGULATIONS AGAINST AI BIAS

Divya Singh, South Africa

12.1 Introduction

Artificial intelligence (AI)[319] describes the broad idea of machines and systems simulating intelligent thought and decision-making previously attributed only to human capability. According to TechTarget, artificial intelligence systems work by ingesting large amounts of labelled training data, analyzing that data for correlations and patterns, and using these patterns to make predictions about future states.[320]

Today, there is common cause that the AI revolution vibrates with the possibilities of untold progress, but the tale does not end here. There is another side to the story. On this other side, there is the meaningful reality of consequential dystopia. Evidence of discrimination and bias, coupled with questions around system accountability and transparency have

[319] Divya Sing is Professor of Law, lawyer, Chief Academic Officer at Stadio Holding of Universities, Cape Town, South Africa. She is Vice-President of the International Board of Globethics.

[320] TechTarget. https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence. (Accessed on 26/06/2024).

seeded grave concerns, but the reality is that the AI systems remain ethically agnostic. "Ethical AI becomes meaningful only when considering how an AI system will impact other human beings."[321]

As we venture into the uncharted territories of AI, it is crucial to focus on responsible AI and AI with integrity. Designers, developers, and deployers in both the public and private sectors—including individuals, businesses, and governments—must carefully consider the potential of machines and machine learning. With society's increasing reliance on AI systems and models, it is imperative to ensure AI integrity, justice, and fair decision-making. This chapter addresses these issues, challenges, and opportunities in three parts: (i) discrimination and bias, human rights, and equality; (ii) discrimination and bias in AI; and (iii) a global synopsis of current legal frameworks addressing AI bias and discrimination.

12.2. Discrimination and bias, human rights and equality

Speaking at the 50th anniversary celebrations of the Universal Declaration of Human Rights, UN Secretary General Kofi Annan exhorted the world to understand human rights as the foundation of human existence and coexistence. Twenty-seven years later his impassioned message remains apposite, as we grapple with the continued vicissitudes of intolerance, discrimination, and injustice. His statement was clear and strong:

> The struggle for universal human rights has always and everywhere been the struggle against all forms of tyranny and injustice …. It is nothing less and nothing different today.
>
> Young friends all over the world, You are the ones who must realise these rights, now and for all time. Their fate and future is in

[321] Artificial intelligence risk management framework, NIST, Sept. 2021. https://www.nist.gov/system/files/documents/2021/09/13/ai-rmf-rfi-0050.pdf

> your hands. Human rights are your rights. Seize them. Defend them. Promote them. Understand them and insist on them. Nourish and enrich them.[322]

Reflecting on that value of equality as a fundamental human right, the *Declaration of the High-level meeting of the [UN] General Assembly on the Rule of Law at National and International Levels* reaffirmed the principles of equality and non-discrimination as part of the foundation of the rule of law. The Declaration further acknowledges the equal application of the rule of law to all states and international organizations and recognised that:

> Respect for and promotion of the rule of law and justice guide all their activities and accord predictability and legitimacy to their actions.
>
> We also recognise that all persons, institutions and entities, public and private, including the state itself, are accountable to just, fair and equitable laws and are entitled without any discrimination to equal protection of the law.[323]

Discrimination is explained in several international instruments, which focus variously either on individual or group experience, or practices and behaviors that give rise to what may be deemed discriminatory conduct. Four selected definitions are presented below:

- *Discrimination is any unfair treatment or arbitrary distinction based on a person's race, sex, religion, nationality, ethnic origin, sexual orientation, disability, age, language, social origin, or other status.* (https://www.un.org)

---

[322] Van der Heijden, B. & Tahzib-Lie, B. (eds). Reflections on the Universal Declaration of Human Rights: A fiftieth anniversary anthology. Martinus Nijhoff Publishers, 1998: Geneva. 18.

[323] Adopted by the General Assembly under Resolution A/RES/67/1* on 30 November 2012.

- *Discrimination occurs when people are treated less favourably than other people are in a comparable situation only because they belong, or are perceived to belong to a certain group or category of people.* (https://www.coe.int)
- *Discrimination occurs when a person is unable to enjoy his or her human rights or other legal rights on an equal basis with others because of an unjustified distinction made in policy, law or treatment.* (https://www.amnesty.org)
- *Discrimination includes any distinction, exclusion or preference made on the basis of race, colour, sex, religion, political opinion national extraction or social origin, which has the effect of nullifying or impairing equality of opportunity or treatment in employment or occupation.* (https://www.ohchr.org)

Discrimination of an individual or a people thus reflects on specific treatment - intended or unintentional - that is different from the treatment given to other persons or populations in the absence of any reasonable and objective justification to support such a decision or behaviour.

Artificial intelligence can impact various human rights individually and/or collectively. However, this chapter focusses particularly on the impact of AI bias and discrimination on the protected rights of equality and non-discrimination. Some country-specific examples are shared that demonstrate the global trend towards legislating to protect against AI bias and promote accountability and transparency of AI systems.

In this context, the understanding of bias provided by Ferrara is extremely useful. He notes:

> Bias is defined as a systematic error in decision-making processes that results in unfair outcomes. In the context of AI, bias can arise from various sources, including data collection. Algorithm design, and human interpretation. Machine learning models, which are a type of AI system, can learn and replicate patterns of bias present

in the data used to train them, resulting in unfair discriminatory outcomes.[324]

The notion of fairness is succinctly used to ground the discussion in this chapter, to simply reflect the absence of bias and discrimination.

12.3 Discrimination and bias in AI

Changes wrought by technology generally and AI and machine learning specifically are unprecedented. There is no gainsaying the significant opportunities to be realized by its utilization and application. Unlike people, machines can continue producing outputs tirelessly and without breaks, they don't become, bored, tired or demand regular working hours, and rather impressively, can consider immense data troves beyond anything known to ordinary human capability. However, at the same time the capacity of AI to violate standards of fairness and infringe upon the fundamental rights of certain groups of people cannot be disregarded. Technology alone will not establish algorithmic accountability, and in this setting, lawmakers cannot ignore the violations to the internationally agreed rights of equality, fair treatment, and non-discrimination. Imagine receiving a university rejection notice because an AI evaluation tool predicted that you showed a higher risk of dropping out, only to find that an AI screening tool built on historic success data had screened you out based on your race.

These and similar other risks have been ever more recorded in the literature, with notable evidence of discriminatory practices and bias especially evident in the operation of AI systems used in predictive

[324] Ferrara, E. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. http://dx.doi.org/10.2139/ssrn.4615421. (Accessed on 25/06/2024).

decision-making models.[325] A survey conducted amongst executives in the U.S. in 2018 underscores the need for concern, with 84% of executives sampled reporting their AI algorithms as only reinforcing existing gender or racial bias.[326] Similarly, data from a 2023 study of business leaders in the U.K. and Ireland shows that only 29% felt confident that "people use AI ethically in business."[327] "Leveraging artificial intelligence to help streamline the hiring process has the potential to amplify biases, because humans are the ones informing the algorithms that make these decisions," notes Douglas.[328]

The first legal challenge in the U.S.A. was settled in *Equal Employment Opportunity Commission (EEOC) v iTutorGroup Inc*. The action by EEOC was founded on the defendants being in breach of the Age Discrimination in Employment, 1967 legislation because its AI hiring programme used 55 years for female applicants and 60 years for male applicants to automatically reject the applications, "resulting in screening out

---

[325] Singh, D. & Singh, A. 2022. AI in Student Recruitment and Selection. 183-204. In Green, E., Singh, D. & Chia, R. (eds). AI Ethics in Higher Education: Good Practice and Guidance for Educators, Learners, and Institutions. Globethics.net, Geneva. When developing an AI model that would filter CVs for employment, Amazon made the honest mistake of training the system with historical data reflecting a predominantly male applicant database. As a result, the output data completely favoured male applicants to the extent that even indirect references to gender (like membership of a women's club), led to the applicant being relegated to a lower level.

[326] *Douglas, E. 2023. AI anti-bias laws: Are algorithms 'inherently' discriminatory? April 13. HRD. https://www.hcamag.com/ca/specialization/employment-law/ai-anti-bias-laws-are-algorithms-inherently-discriminatory/. (Accessed on 24/06/2024).*

[327] A. Miller. 2024. 7 essential practices for ensuring ethical and bias-free AI in business analytics. February 21. Analytics. *https://pubsonline.informs.org/do/10.1287/LYTX.2024.01.12/full/*. (Accessed on 29/5/2024).

[328] Douglas above fn 7.

over 200 applicants because of their age." The defendants accepted liability, agreed to pay $365 000 (USD) to the group of affected applicants, adopt anti-discrimination policies, and conduct training to ensure compliance with equal employment opportunity laws.[329]

The pending outcome in *Mobley v Workday Inc* is eagerly anticipated for the precedent it could set on the legal implications of using AI to automate hiring and other employment functions. The case challenges the defendant's algorithm-based applicant screening tool for bias based on race, age, and disability, further claiming that the training data did not account for existing discrimination in the sample. The defendant argued that the existing legislation holds employers liable for discrimination when it is caused by a contractor or other third party, but there is no precedent extending liability to a third-party business with no control over hiring decisions. The defendants thus argued that as a software vendor they were not covered by the existing laws. The judge expressed concern that shielding software vendors from the national anti-discrimination laws would leave victims of discrimination with no legal recourse and confirmed the more-favored broad application of Title VII of the Civil Rights Act, 1964 "in order to carry out the law's purpose of eradicating discrimination. That includes applying the law to entities that can affect a worker's access to job opportunities."[330] Quoting the presiding judge in the matter:

---

[329] Kempe, L. 2024. *Navigating the AI employment bias: Legal compliance guidelines and strategies.* 3. https://www.americanbar.org/groups/business_law/resources/business-law-today/. (Accessed on 23/06/2024).

[330] Wiessmer, D. 2024. Workday urges judge to toss bias class action over AI hiring software. May 14. *Reuters*. https://www.reuters.com/legal/transactional/workday-urges-judge-toss-bias-class-action-over-ai-hiring-software-2024-05-14/. (Accessed on 23/06/2024).

> What troubles me about Workday's interpretation is the concept that the employer would not be liable for intentional discrimination unless they knew that the software was (biased).[331]

Entering the fray, the U.S. Equal Employment Opportunity Commission, the national body responsible for enforcing Title VII, urged that Title VII was applicable to Workday Inc because its software decides which applicants will be considered for a job at all.[332] Judgment in the matter is awaited.

Explaining the notion of AI bias, IBM describes it as: *[AI systems that produce] biased results due to human biases that skew the original training data or AI algorithm – leading to distorted outputs and potentially harmful outcomes*.[333]

Kempe refers to: AI systems that produce biased results that reflect and perpetuate human biases within a society, including historical and current social inequality.[334]

For the purposes of this discussion, algorithmic bias is applied in the context of predictions or outputs from an AI system, where those predictions or outputs exhibit erroneous or unjustified differential treatment between two groups.[335] Thus, algorithmic bias will constitute, or may lead to unlawful discrimination where it is not reasonably possible to show legal justification for the difference in outcome.

AI bias (also referred to as machine learning bias) may be the result of training data or/and programming errors where the flaw arises from

---

[331] Wiessmer above fn 11.

[332] Wiessmer above fn 11.

[333] IBM. 2023. *What is AI bias.* December 22. https://www.ibm.com/topics/ai-bias. (Accessed on 23/06/2024).

[334] Kempe above fn 10.

[335] Australian Human Rights Commission. 2020. Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias. Technical Paper. 13. ISBN: 978-1-925917-27-7.

training data: AI systems build their predictive decision-making capabilities on training data ("input data"). When the training data is skewed to under- or over-represent one or other group or a specific factor or collection of factors, or if the training data reflects historic or current social biases, the algorithms trained on them will mimic these biases, reflecting them in their predictions and giving rise to biased results ("output data"). Programming errors on the other hand arise as a factor of the developer's input, where a programmer deliberately or unconsciously emphasizes considerations or particular features when programming the algorithmic model. Examples of the latter are often the result of the subjective opinions of the programmer. Bias can also result from algorithms themselves either amplifying existing biases in the data or, due to certain design choices, creating their own bias even if the data itself is not biased.

The outcomes of these biased algorithms can then be fed into real world systems and affect users' decisions, which will result in more biased data for training future algorithms.[336] Whatever the genesis, the fact is that biased algorithm outcomes can impact user experience, generating a feedback loop between data, algorithm and users that can perpetuate and even amplify existing sources of bias.

"In theory, AI is objective – but in reality, AI systems are informed by human intelligence, which is of course far from perfect."[337] However, the potential for AI bias and discrimination to go unchallenged is often a consequence of what has been described as 'AI confidence'. "In direct human communication," notes Leffer "subtle cues of uncertainty are

---

[336] Mehrabi, N. et al. 2019. *A survey on bias and fairness in machine learning.* chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/1908.09635. (Accessed on 24/06/2024).

[337] Tait, E.J., Galvan, H.C. & Linas, J.M. 2019. Proposed Algorithmic Accountability Act targets bias in artificial intelligence. June. *Insights.* https://www.jonesday.com/en/insights/2019/06/proposed-algorithmic-accountability-act. (Accessed on 29/05/2024).

important for how we understand and contextualize information."[338] However, with machines, models, and systems this is not so. Further, there is a general perception that algorithms are inherently neutral and, therefore, must be objective. Also, suggests Lemmer, people attribute greater authority to AI systems "because algorithms are often marketed as drawing on the sum of all human knowledge."[339] In similar vein but perhaps more caustically, Vicente and Matute state:

> Human trust in automation influences the tendency to over-accept algorithmic outcomes even when they are noticeably wrong. This means that humans are not only willing to rely on AI because they are "cognitive misers", that take mental shortcuts when making decisions, but also because they perceive artificial intelligence to be trustworthy. It has been suggested that trust could induce compliance with AI advice due to an authority effect.[340]

All of this, however, misses the central fact that all AI models are built on rules created by human beings and are consequently susceptible to developer biases. As noted by the Panel for the Future of Science and Technology:

> AI enables automated decision-making even in domains that require complex choices, based on multiple factors and non-predefined criteria. In many cases, automated predictions and decisions are not only cheaper, but also more precise and impartial than human ones, as AI systems can avoid the typical fallacies of human

---

[338] Leffer, L. 2023. *Humans absorb bias from AI – And keep it after they stop using the algorithm.* October 26. Https://www.scientificamerican.com/article/humans-absorb-bias-from-ai-and-keep-it-after-they-stop-using-the-algorithm/. (Accessed on 29/5/2024).

[339] Leffer above fn 19.

[340] Vicente, L. & Matute, H. 2023. Humans inherit artificial intelligence biases. 13(15737). *Nature.* https://doi.org/10.1038/s41598-023-42384-8. (Accessed on 29/5/2024).

> psychology and can be subject to rigorous controls. However, algorithmic decisions may also be mistaken or discriminatory, reproducing human biases and introducing new ones. Even when automated assessments of individuals are fair and accurate, they are not unproblematic: they may negatively affect the individuals concerned, who are subject to pervasive surveillance, persistent evaluation, insistent influence, and possible manipulation.[341]

Perhaps, however, the public tide is turning. The 2024 *Artificial Index Annual Report* conducted by Stanford University records that in the last year, 66% of respondents (an increase from 60%) believed that AI "will dramatically affect their lives" in the next 3-5 years, but 52% (up from 38% in the last year) expressed reservations towards AI products and services. Only 56% of the respondents trusted AI "to not discriminate or show bias towards any group of people". Similarly, only 52% of the survey trusted companies using AI as much as they trusted other companies.[342]

AI systems used for automated decision-making use data and machine learning algorithms to train mathematical models for the purposes of prediction and/or decision-making. The training data includes historical data and information including existing decisions and informing factors. The AI system is trained to identify patterns within its datasets as well as common features associated with particular types of decisions. Fostering AI integrity would thus have to start from conception of the system, identification of appropriate data to inform the system, collection of data, and

---

[341] Panel for the Future of Science and Technology. 2020. *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence*. European Parliamentary Research Service [online]. i. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf. (Accessed on 20/06/2020).

[342] Stanford University HAI. 2024. *2024 AI Index Report*. Chapter 9. Available from https://aiindex.stanford.edu/about/. (Accessed on 26/06/2024).

design of the system. In this phase, data quality and accuracy is key, and may be promoted by ensuring (i) correctness, (ii) diversity and inclusivity, and (iii) impartiality and representivity of information used. Secondly, developers must implement audits and equity assessments to ensure that inadvertent bias has not crept into the system.

Sandbox engagement and thorough and exacting testing of outcomes, with a lens on multi-stakeholder impacts will ensure that systems are robust for deployment. Further understanding that data in the larger system has the potential to act unpredictably and amplify itself as the system is continually gathering new data and updating how it behaves,[343] AI models – and especially those used for predictive decision-making - must continue to be rigorously monitored with iterative testing even after original implementation checking for programme degeneration, discrimination, and bias. Miller's[344] simple check list highlights the following checks:

- Are all groups accurately [and proportionately] represented in the data?
- Are the models appropriate for the desired use cases?
- Has someone (and preferably more than one person) checked to ensure that there are no skewed outcomes? Miller is a strong proponent of maintaining the human oversight and she quotes a 2023 study of business leaders in the U.K. and Ireland where 93% of the respondents also favored maintaining humans' involvement in AI-based decision-making.
- Have algorithmic developers defined the key metrics?
- Will the algorithms include a fairness regulator to reduce bias?

[343] Joshi, N. 2022. *How to stop AI from going rogue?* November 4. https://www.allerin.com/blog/how-to-stop-ai-from-going-rogue. (Accessed on 28/05/2024).

[344] Miller above fn 8.

Human rights must be the underlying lens across the entire value chain and assured at all stages of the design, development and deployment of the technology. This is in line with the *Guiding Principles on Business and Human Rights* published by the UN Office of the High Commissioner which emphasizes the proactive identification of human rights risks or impacts of their activities and relationships.[345] Applying the guidelines as a principle of best practice would thus impose a direct responsibility on businesses designing AI as well as organizations deploying the technology to ensure the lawful and responsible use of the tools that are placed in the public domain.

Accountability for AI biases with discriminatory outcomes and consequential unfair and unjust decision-making raises interesting questions in law; and as the capabilities of AI systems increase at a record pace, the need for policy guidelines at least and responsive legal frameworks at best governing their safe and responsible implementation becomes crucial.[346]

Further reflecting on the challenges of accountability, two important considerations bear mention – firstly, an awareness that machine learning algorithms analyze huge datasets, searching for patterns and extrapolating beyond training-set examples; and secondly, biased training data can remain hidden as software developers unintentionally embed bias into the design and operation of the system.[347] For example, the decisions made by developers when selecting algorithms and modelling approaches, the

---

[345] Australian Human Rights Commission above fn 16, 13.

[346] Zartis Team. n.d. *AI Legislation in the US: AI Regulation Review.* 4. https://www.zartis.com/ai-regulation-review-ai-regulation-in-the-us/. (Accessed on 29/05/2024).

[347] Chiappetta, A. 2023. Navigating the AI frontier: European parliamentary insights on bias and regulation, preceding the AI Act. 12(4). *Internet Policy Review.* 4. https://policyreview.info/articles/analysis/navigating-ai-frontier-european-parliamentary-insights. (Accessed on 29/05/2024).

definitions of success applied in a particular context, or the selection of traits to be included in a model can introduce bias into AI.[348] Further down the line, the biased data can result in reproduction and amplification of those biases in predictive AI models projecting historical patterns into the future and perpetuating a cycle of unfair discrimination against the already marginalized communities. In this milieu, the AI developers' own limited understanding of the AI systems internal mechanisms will complicate issues of accountability. Even more disconcerting is the amplification of bias.

In a study conducted by Vicente and Matute looking at AI recommendations used in the health care sector, they observed that erroneous and biased recommendations made by the AI systems had the potential adversely influence human behavior in the long term because humans perpetuated the same biases displayed by the AI even after the engagement with the biased model and in response to new stimuli and engagements.[349] The effect is that the bias has the potential to become integrated into the belief system and embedded in how issues are perceived and decisions are made.

"If left unchecked, AI bias has the potential to trigger significant societal ramifications, such as engendering less diverse workforces, increasing incarceration rates, or exacerbating income and wealth disparities," stress Sher and Benchlouch[350] and the problem will just increase if left unchecked. The newly applied Foundation Model Transparency Index reported in the *2024 AI Index Annual Report* shows that AI

---

[348] activeMind.Legal n.d. *Bias in artificial intelligence: Risks and solutions.* https://www.activemind.legal/guides/bias-ai/. (Accessed on 29/05/2024).

[349] Vicente et al above fn 21.

[350] Sher, G. & Benchlouch, A. 2023. Unmasking AI bias: a collaborative effort. July 21. *Reuters.* https://www.reuters.com/legal/legalindustry/unmasking-ai-bias-collaborative-effort-2023-07-21/. (Accessed on 29/05/2024).

developers "lack transparency regarding disclosure of training data and methodologies,"[351] rendering regulation and legislation in this space an urgent matter. Emphasizing the opaque nature of AI models and systems Wachter et al explain the problem as follows:

> A fundamental challenge arises from the non-decomposable nature of complex modern machine learning or AI systems. While indirect discrimination has been detected in complex systems such as, for example, governmental provision housing in Hungary, this has in no small part been possible because of the judiciary to decompose these large systems into isolated components or a collection of housing policies, and to evaluate whether each such policy satisfies non-discrimination law. This decomposition has been an incredibly effective tool that has allowed judges to bring 'common sense reasoning' to bear on complex and nuanced systems, and also strengthens the powers of statistical reasoning, making the choice of test less significant. However, attempting to understand discrimination caused by complex algorithmic systems that are not similarly decomposable reveals the limitations of common-sense reasoning.[352]

Without readily available information on the systems composition, claimants will always have trouble proving automated discrimination. Claimants will often not have the information regarding a system's optimization conditions or decision rules, and thus be unaware of the reach and definition of the contested rule that led to perceived disparity. Similarly, without information regarding the scope of the system and the

---

[351] Stanford University HAI above fn 23.

[352] Wachter, S., Mittelstadt, B. & Russell, C. 2020. *Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI.* Available at https://www.are.na/block/10663562. (Accessed on 26/06/2024).

outputs or decisions received by other individuals, claimants will have difficulty defining a legitimate comparator group.

12.4 Assuring equality, and legislating against AI discrimination and bias: A country-specific synopsis of laws and legislation

12.4.1 Africa

In 2021, Africa's AI Blueprint was launched, emphasizing the importance of ethical considerations in AI adoption which must be addressed through policy and regulatory approaches. In the same year, the African Commission on Human and Peoples' Rights adopted Resolution 473 which outlines the need to address the implications for human rights of AI, robotics and other new and emerging technologies in Africa. However, while a core set of general principles has emerged for the Continent, they have not yet been operationalized through legal regulations.[353] In July 2024, the African Union also published the Continental Artificial Intelligence Strategy: Harnessing AI for Africa's development and prosperity.

12.4.2 Brazil

In December 2022 Brazil considered draft regulations for responsible AI governance which prioritized (i) guaranteeing the rights of people affected by AI; (ii) classification of risk levels; and (iii) predicting

[353] Mpedi, L.G. 2023. Opinion: Prepare for AI: As we navigate the era of artificial intelligence, South Africa must choose a regulatory path. August 18. *UJ News.* https://news.uj.ac.za/news/prepare-for-ai-as-we-navigate-the-era-of-artificial-intelligence-south-africa-must-choose-a-regulatory-path/. (Accessed on 29/05/2024).

governance measures. However, Bill No 2,338/2023 is still pending before the Parliament and no law has been approved.[354]

12.4.3 Canada

In 2022 the federal government of Canada released the *Digital Charter Implementation Act* which includes Canada's first piece of legislation to mitigate bias in AI. It was followed by the *Artificial Intelligence and Data Act* which imposes an onus on developers specifically to responsibly ensure the mitigation of risk and bias, while also including prohibitions on harmful use and requirements of public disclosure.[355]

12.4.4 China

In 2017 the *Next Generation Artificial Intelligence Development Plan* was released setting guidelines for the use of specific AI applications; and in 2021 China published a *New Generation Artificial Intelligence Code of Ethics* which prescribes:

> Improvement of human well-being: AI systems should follow common values, respect human rights and the fundamental interests of society, promote harmony, improve livelihoods, and take a sustainable approach to economic, social, and ecological development.
>
> Promotion of fairness and justice: AI systems should be inclusive and effectively protect the rights and interests of those that interact with the system while sharing the benefits of AI across society to promote fairness, justice, and equal opportunities. Vulnerable

---

[354] Actian Corporation. 2023. *The global race to regulate bias in AI.* July 11. https://www.actian.com/blog/data-management/bias-artificial-intelligence-big-data/. (Accessed on 29/05/2024); 2024. *AI Watch: Global regulatory tracker – Brazil.* May 13. Available at https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-brazil. (Accessed on 28/06/2024).

[355] Actian above fn 35.

> groups should be respected, and accommodations should be provided where necessary…[356]

The *Internet Information Service Algorithmic Recommendation Management Provisions* came into effect in 2022. Among many things, the regulation mandates increased transparency and audits of recommendation algorithms. To facilitate its commitment to transparency and fairness, the regulations establish an algorithm registry responsible for understanding how all algorithms work and ensuring that application is within acceptable parameters.[357] In 2023 the *Interim Measure for Generative Artificial Intelligence Service Management* was passed and is geared at balancing innovation and legal governance. Key principles encompass the obligation to respect the rights of others and not "promote discrimination" which specifically includes taking measures to prevent discrimination on ethnicity, belief, nationality, religion, gender, age, occupation and health resulting from generative AI. Further, explicitly stated is the need to take measures to "improve transparency, accuracy, and reliability."[358]

The *Personal Information Protection Law* (2021) also sets out rules regarding automated decision-making promoting transparency, fairness and justice of the automated results; and prohibiting "unreasonable differential treatment of individuals based on automated decision-making." Where an individual's rights are "significantly affected" s/he may request an explanation of how the AI-generated outcome will be used and can

---

[356] Kachra, A. 2024. Making sense of China's AI regulations. February 12. *Holistic AI.* https://www.holisticai.com/blog/china-ai-regulation. (Accessed on 24/06/2024).

[357] Kachra above fn 37; Chen, G. 2024. Balancing innovation and regulation: Comparing China's AI Regulations with the EU Act. April 11. *AWA Point.* https://awapoint.com/balancing-innovation-and-regulation-comparing-chinas-ai-regulations-with-the-eu-ai-act/. (Accessed on 29/05/2024).

[358] Kachra above fn 37.

also "prohibit the handler from making decisions based solely on its use." Provision is also made for mandatory impact assessments, with clear guidelines on what should be evaluated. The rules apply to natural and juristic persons in-country as well as outside China under specified conditions. Another key piece of legislation is the *Administrative Measures for Generative Artificial Intelligence Services*, 2023 which explicitly mandates the use of truthful, accurate data, free of discriminatory algorithms, prioritizing prevention as the critical forerunner of responsible AI.[359]

12.4.5 European Union

The AI Act passed by the European Parliament has been variously described as a pioneering model for other countries,[360] and a landmark attempt to regulate AI with a prioritization of what is classified as "high risk AI" which, states Kaplan, includes systems that have tangible and potentially life-changing impact on peoples' lives.[361] As pointed out by Boda et al not all AI can be regarded in the same way when covering AI risk management. The Act therefore creates four risk-based categories namely, *unacceptable risk* which is banned; *high risk* which is subject to strict regulatory requirements; *limited risk*; and *minimal risk*.[362] Classifying algorithmic technologies into categories of risk, allows the EU to

[359] Actian above fn 35.

[360] Chiappetta above fn 28, 12.

[361] Kaplan, M. 2021. Can you legislate AI? April 28. *CallMiner.* https://callminer.com/blog/can-you-legislate-ai. (Accessed on 29/05/2024).

[362] Boda, R., Gunning, E. & Ntuli, L. 2024. The EU Act passes: Should South Africa follow suit and regulate artificial intelligence. March 19. *ENSight*. https://www.ensafrica.com/news/detail/8261/the-eu-ai-act-passes-should-south-africa-foll#:~:text=. (Accessed on 29/05/2024).

effectively limit high risk technologies and outright ban software they rank "unacceptable", notes Riccardi.[363]

Specifically, the AI Act limits negative effects and prejudice in areas that pose risks to health, safety, and fundamental rights of persons. With reference to that last-mentioned, section 31 provides:

AI systems providing social scoring of natural persons by public *or private actors* may lead to discriminatory outcomes and the exclusion of certain groups. They may violate the right to dignity and non-discrimination and the values of equality and justice. Such AI systems evaluate or classify *natural persons or groups thereof* on the basis of *multiple data points related to* their social behaviour in multiple contexts or known*, inferred* or predicted personal or personality characteristics *over certain periods of time*. The social score obtained from such AI systems may lead to the detrimental or unfavourable treatment of natural persons or whole groups thereof in social contexts, which are unrelated to the context in which the data was originally generated or collected or to a detrimental treatment that is disproportionate or unjustified to the gravity of their social behaviour. *AI systems entailing such unacceptable scoring practices and leading to such detrimental or unfavourable outcomes* should be therefore prohibited.

Title II of the Act uses a risk-based approach to prohibit AI that creates an "unacceptable risk". AI that violates fundamental rights is specifically mentioned.[364] A key provision of the Act insofar as it seeks to curtail bias and perpetuating discrimination is the emphasis in regulating input data on the basis of which an AI system is trained to produce an output - Articles 13 and 26. As pointed out by Verdict unbalanced data is one of

[363] Riccardi, G. 2023. America's first law regulating AI bias takes effect this week. July 3. *Quartz.* https://qz.com/americas-first-law-regulating-ai-bias-in-hiring-takes-e-1850602243. (Accessed on 29/05/2024).

[364] Kaplan above fn 42, 3.

the main reasons for misrepresentation in AI models.[365] Given the genesis of machine learning algorithms, models can inherently only be as good as the data used to train them. Training data that is compromised by systematic bias towards certain individuals or groups of individuals will, with certainty, result in the algorithms not only learning and repeating the biases, but alco consolidating and reinforcing them in their output processes. Datasets need to be representative and balanced to produce a balanced output. "Regulating input data will thus improve AI explainability and concomitantly reduce algorithmic bias."[366]

Kaplan also suggests that a further way to regulate AI would be to regulate the types of outcomes that are useable in society. It is trite that with predictive AI systems data is key in building the model: however, it is a reality that data often contains many biases, with the further potential for self-generated bias. Kaplan therefore argues for regulations that will set accuracy thresholds for use and supports the existence of "standards for evaluating how a model's performance changes over time and making sure that people can intervene as necessary."[367] Verdict concurs with the importance of system monitoring as a crucial factor of continued system integrity and accuracy.[368] This is covered by Article 15 of the Act which deals with system accuracy and robustness and regulates the risk of model changes:

> High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations ('feedback loops'),

---

[365] Verdict. 2021. *Landmark AI legislation could tackle algorithmic bias.* April 28. 1. https://www.verdict.co.uk/ethical-ai-regulation-eu/. (Accessed on 29/05/2024).

[366] Verdict above fn 46, 2.

[367] Kaplan above fn 42, 3.

[368] Verdict above fn 46, 2.

and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures.

Summarizing the ethos of the Act, Chiappetta states, "The AI Act seeks to achieve an equitable and EU-wide standard for AI regulation while finding the balance between ethics, innovation and safeguarding rights."[369] The regulations are explicit in ensuring and promoting transparency and accountability, while also making specific provision for responsible process management as well as monitoring and assessment of systems placed in society. On the other hand, the drafters of the legislation were conscious not to stifle innovation balancing the protection of human rights with the promotion of "things like building sandboxes or controlled development areas that fit regulatory standards, in which creators can test their AI systems."[370]

The AI Act provides protection for all EU citizens, and any company operating within the EU or utilising EU citizens' data would be compelled to comply with its provisions. The impact is that companies worldwide doing business in the EU will need to adhere to these standards regardless of the regulations (or lack thereof) in the home country creating, according to Verdict "the foundations for a global standard for ethical AI."[371]

12.4.6 Japan

The release of the *Social Principles of Human-Human-Centric AI*, 2019 by the national Integrated Innovation Strategy Promotion Council established seven social principles including human-centricity, fairness, accountability and transparency, and innovation to regulate the public and private use of AI in Japan.

---

[369] Chiappetta above fn 28, 2.

[370] Section 25 of the AI Act, 2023; Kaplan above fn 42, 4.

[371] Verdict above fn 46, 3; also, Kaplan above fn 42, 5.

12.4.7 South Africa

South Africa has no explicit legal framework for AI and remains largely unregulated. However, existing legislation like the *Protection of Personal Information Act*, 2013 (POPIA) provides some limited protection regarding AI bias and discrimination.

POPIA requires all entities using AI or machine learning for purposes of automated decision-making to ensure that the affected parties are adequately informed of the intention. Section 71 of POPIA deals specifically with the question of automated decision-making. Sub-section (1) provides that a data subject may not be subject to a decision which results in legal consequences for them, or which affects them to a substantial degree, which is based solely on the automated processing of personal information intended to provide a profile of that person. Sub-section (2) sets out certain exceptions to the general prohibition, such as whether the decision is in connection with the conclusion or execution of a contract, and appropriate measures are in place to protect the data subject's legitimate interests. In this regard, "appropriate measures" require that the data subject has an opportunity to make representations about the decision and is provided with sufficient information about the underlying logic of the automated processing of the information to make such representations. Singh et al suggest that the insertion of this provision evinces, at least, some understanding from the legislators of the potential for risk attendant upon automated decision-making, and the broader implications that this may have on affected persons.[372]

12.4.8 United States of America

Unlike the centralized EU Regulations, the U.S. has no overarching federal legislation focusing on protecting people from the potential harms of AI and other automated systems. Rather, the legal landscape remains

[372] Singh et al above fn 6.

a patchwork of sector-specific regulations, Agency guidelines (including Agencies such as the National Institute of Standards and Technology's (NIST) "Artificial Intelligence Risk Management Framework"), Executive Orders, and State legislation.[373]

The *Blueprint for an AI Bill of Rights* was published by the Biden administration in October 2022 and outlines the five key principles - safety, fairness, accountability, transparency, and privacy - that should inform AI development and use. While having no legal force and effect, Zartis suggests that its value lies in the overarching commitment that is made "to ensuring ethical and responsible AI development in the US."[374]

In October 2023 the President issued Executive Order 14110 titled "Safe, Secure and Trustworthy Development and Use of Artificial Intelligence" which emphasized the need for responsible AI development. It focuses specifically on issues of safety, security, bias and accountability. According to Zartis the Order does not impose explicit regulations, but outlines guiding principles focused on federal agencies directing them to, at least, develop specific plans for their responsible use of AI.[375]

Friedler et al confirm the importance of sector-specific legislation agreeing that existing state agencies are best positioned to understand the role and impact of algorithmic systems in their domains and provide the necessary oversight. However, they remind us that anticipated benefits will only be realized if the various state agencies have the requisite technical expertise to effectively oversee algorithmic systems.[376]

---

[373] Zartis above fn 27; Friedler, S., Venkatasubramanian, S. & Engler, A. 2023. How California and other states are tackling AI legislation. March 22. *Brookings*. 3. https://www.brookings.edu/articles/how-california-and-other-states-are tackling-ai-legislation/. (Accessed on 29/05/2024).

[374] Zartis above fn 27, 5.

[375] Zartis above fn 27, 3.

[376] Friedler et al above fn 54, 3.

Under federal sector specific legislation, the Federal Trade Commission (FTC) has regulated the equitable use of AI under its *Fair Credit Reporting Act.* The Act makes provision for any vendor collecting data for the purpose of automating decision-making about an applicant's eligibility for credit, employment, housing, or similar benefits to be potentially a "consumer reporting agency". According to Mithal this triggers duties for companies that use the vendor's services including the requirement to provide an adverse action notice to the applicant. Explaining further, Mithal uses the example of an employer who purchase AI-based scores for assessing whether an applicant will be a good employee: if the employer denies the applicant a job based on the scores, it must provide the applicant with "an adverse action notice" informing the applicant that s/he may access the underlying information from the vendor and correct it if it proves to be false.[377]

Through its powers under the *Federal Trade Commission Act*, its effect in commercial practices, makes the use of algorithms that discriminate against protected classes a possibly unfair practice. The Act also prohibits unfair or deceptive practices so that if a party "makes false or unsubstantiated claims about lack of bias in its algorithm, that could be a deceptive practice."[378] The third relevant piece of legislation from the FTC is the *Equal Credit Opportunity Act* explicitly banning discrimination in credit decision-making based on color, race, religion, nationality, sex, marital status, age, or the use of public assistance.[379]

---

[377] Mithal, M. 2022. *Legal requirements for mitigating bias in AI systems.* February 15. https://www.jdsupra.com/legalnews/legal-requirements-for-mitigating-bias-3221861/. (Accessed on 29/05/2024).

[378] Mithal above fn 58.

[379] Actian above fn 35.

In May 2023 the EEOC issued guidelines on how to measure adverse impact when AI tools are used for employment selection.[380] Failure to adopt a less discriminatory tool renders the employer liable for discriminatory conduct; and further, renders the employer liable for the actions of external developers /administrators of any algorithmic decision-making tool where the vendor has acted on behalf of the employer. The Act clarifies that employers cannot oust liability by relying on the vendor's assessment of the tool's impact.[381]

Many states in the U.S.A. have developed specific laws or frameworks seeking to regulate AI use, limit bias and discrimination. All state legislation further seeks to create stronger protections through accountability and transparency mechanisms. However, notwithstanding important areas of consensus, there is no single model that serves all states. This is especially noticeable in the ambit of state oversight. Friedler et al explain, "So far, state legislators have reached different decisions about whether to limit their oversight to government uses of these systems or whether to consider other entities within the state, especially the commercial use of algorithms."[382]

[380] U.S. EEOC. May 18, 2023. "Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964".

[381] Kempe above fn 10, 7.

[382] Friedler et al above fn 54.

12.5 State regulatory and legal frameworks specifically aimed at preventing AI bias and discrimination[383]

California, 2023: *California Privacy Rights Act*: governs profiling and automated decision-making, defining "profiling" as *any form of automated processing of personal information … to evaluate certain personal aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behavior, location, or movements.* The Act gives consumers opt-out rights with respect to businesses' use of automated decision-making technology. Additionally, when AI systems are used for decision-making, the Act places an onus on employers and business to ensure that affected parties know when and how the system is being used. Both parties developing and deploying AI systems have the responsibility to ensure compliance with the legislated norms governing AI systems and will be held accountable for non-compliance. The requirement for impact assessments (introduced in 2024), however, failed before the state legislature. California further specifically requires criminal justice agencies using AI-powered pre-trial risk assessment tools to analyze whether the systems produce bias based race, gender pr ethnicity.[384]

Colorado, 2021: *Protecting Consumers from Unfair Discrimination in Insurance Practices*: safeguards consumers from unfair discrimination by insurance providers using rate-setting mechanisms. The law is sector limited and prohibits insurers from using consumer data gathered by AI systems and predictive models in insurance decisions in a way that

---

[383] BCLP Client Intelligent. 2023. *US State-by-State AI Legislation Snapshot.* https://www.bclplaw.com/en-US/events-insights-news/2023-state-by-state-artificial-intelligence-legislation-snapshot.html (Accessed on 29/05/2024).

[384] Wright, R. *Artificial intelligence in the States: Emerging legislation*. 2023. December 6. https://www.csg.org/2023/12/06/artificial-intelligence-in-the-states-emerging-legislation/. (Accessed on 24/06/2024).

unfairly discriminates against any persons based on "race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression.[385] The *Colorado Artificial Intelligence (AI) Act*, 2024 has been described as the first piece of comprehensive AI regulation to protect consumers passed in the U.S.[386] The Act specifically targets AI discrimination and emphasizes transparency especially in high-risk AI bias areas including cases where Coloradans have applied for a job, a financial service, an educational programme, or home.[387] The legislation directs developers and deployers to take reasonable care to avoid discrimination, implement adequate risk management policies to prevent discrimination, conduct impact assessments of their AI models to ensure that they are free from discrimination, and report to the appropriate authorities on the continued health of the model. Acknowledging the black box nature of most AI systems, developers are obligated to disclose information about their systems and specifically how they've tested for bias throughout the process.[388] The regulations make it obligatory that affected parties are informed that they're dealing with a machine, and the purpose for which the technology-driven system will be used.

---

[385] Wright above fn 65.

[386] Birkeland, B. 2024. In writing the country's most sweeping AI law, Colorado focused on fairness, preventing bias. June 22. *CPR News.* https://www.npr.org/2024/06/22/nx-s1-4996582/artificial-intelligence-law-against-discrimination-hiring-colorado. (Accessed on 24/06/2024).

[387] Klamann, S. 2024. Colorado's first-in-nation AI regulations are "a major shift" experts say, as tech industry pushes back. May 24. *Denver Post.* https://www.denverpost.com/2024/05/24/colorado-artifical-intelligence-ai-law-regulations-tech-congress-discrimination/; DPO Centre. 2024. *Colorado becomes first state to pass AI legislation*. May 21. https://www.dpocentre.com/colorado-becomes-first-state-to-pass-ai-legislation/. (Accessed on 24/06/2024).

[388] Klamann above fn 68.

Illinois, 2020: *Artificial Intelligence Video Interview Act* (820 ILCS 42/1): imposes obligations on employers if they conduct video interviews and use AI analysis of the videos in their evaluation process (notifying applicants of the AI's role- when and how it is being used-).

Maryland, 2020: *Facial Recognition Technology Law*: prohibits employers from using certain facial recognition services – many of which use AI processes – during job interviews unless the applicant consents.

New York City, 2021: *Automated Employment Decision Tools (AEDT) Law*: prohibits employers and employment agencies from using AEDT to assesses candidates for hiring or promotion unless there has been an independent bias audit of the AEDT, and applicants are notified of use. This is the first law in the U.S. requiring employers to conduct bias audits of AI-enabled tools used for employment decisions.[389] Employers using AI technology as part of their hiring process must perform an annual audit of their technology to evaluate bias – intentional or otherwise – built into these systems. The audit results must be published on the corporate website. "By mandating comprehensive anti-bias audits, the law empowers employees to challenge their biases and create a more inclusive and fair working environment."[390] Bias is evaluated with an

[389] In District of Columbia the *Stop Discrimination by Algorithms Act* (2023) which sought to prohibit all organisations from using algorithms that make decisions based on protected personal traits was not approved. Similarly, in Illinois (2023) the amendment to the *Human Rights Act* providing that employers using predictive data analytics in their employment decisions should not consider the applicant's race, or zip code when used as a proxy for race, to reject an applicant failed. Similar legislation (2022) regulating the use of automated tools in hiring decisions and minimizing discrimination in employment was not approved by New Jersey.

[390] Kestenbaum, J. 2023. NYC's new AI bias law broadly impacts hiring and requires audits. July 5. *Bloomberg Law.* https://news.bloomberglaw.com/us-law-week/nycs-new-ai-bias-law-broadly-impacts-hiring-and-requires-audits. (Accessed on 29/05/2024).

"impact ratio" measuring the effect of the technology on hiring across legally protected groups "defined by race, ethnicity, and gender – although notably," remarks Riccardi, "not people with disabilities or older workers."[391] The legislation also provides for candidates or employees to request alternative evaluation processes as an accommodation. Contraventions by employers of the regulated provisions will expose employers to the sanction of a fine. - Two further pieces of progressive legislation in this area pending before the New York state legislature include (1) laws making it unlawful for a landlord to implement or use an automated decision tool unless it (i) conducts a disparate impact analysis to assess actual impact of the tool and publishes the assessment, and (ii) notifies applicants of use; and (2) the *New York Artificial Intelligence Bill of Rights.* The proposal affords all residents of New York rights and protections including the right to safe and effective systems, protection against algorithmic discrimination and abusive data practices, and the right to know when an automated system is being used, and how it contributed to the outcome. If passed, all affected parties have the unchallenged right to opt out of automated decision making with the correlative right to work with a human in place of an automated system.

Vermont: the draft legislation (2024) requires that developers and deployers of "high risk AI systems" use reasonable care to avoid any risk of algorithmic discrimination that is a reasonably foreseeable consequence of utilizing the system. Specific obligations and minimum standards for developers and deployers are stipulated in the bill linked to the *Artificial Intelligence Risk Management Framework* published by NIST.

The State of Georgia introduced the *South Carolina Age-Appropriate Design Code Act* in 2024, prohibiting businesses from profiling children under the age of 18 unless able to demonstrate appropriate safeguards to protect the best interests of the child and either that profiling is

[391] Riccardi above fn 44.

necessary, or there is a compelling reason that profiling will be in the child's best interests.

Several states have also proposed or enacted privacy and data protection legislation giving consumers a right to opt-out of the processing of their personal data for purposes of profiling in furtherance of decisions that produce legal or similarly significant effects.[392]

At the federal level, several pieces of legislation to eliminate algorithm and machine learning bias and discrimination are currently under consideration by Congress including the *Algorithmic Accountability Act* which seeks to establish transparency and accountability for algorithmic decision-making, emphasizing the importance of assuring the lawful and ethical use of AI. The Act obliges implementing organizations to understand the workings of the algorithms being employed before use and to take responsibility for "reasonably addressing in a timely manner" any identified biases or security issues.[393] The second is *The Algorithmic Justice Leage Act* which is a response to the acknowledged prevalence of algorithmic bias and discrimination in areas such as employment, housing, and criminal justice.

Summarizing the purpose of the legislation, Zartis notes: *[The] proposal seeks to enact targeted measures to address these issues … [and] the Act aims to develop comprehensive strategies for identifying and mitigating algorithmic bias. It calls for the establishment of mechanisms to hold accountable entities accountable for the adverse impacts of biased algorithms, thereby safeguarding against systemic injustices perpetuated by AI-driven systems*.[394]

---

[392] Interesting, similar regulation allowing individuals to opt-out of processing of the personal data for the purpose of future profiling failed at the West Virginia state legislature.

[393] Tait et al above fn 18.

[394] Zartis above fn 27, 4.

The *Eliminating Bias in Algorithmic Systems Act* was introduced in December 2023. It concentrates on accountability, mandating the establishment of civil rights offices by all federal agencies using, funding, or overseeing AI which will focus on combating AI bias and discrimination".[395] Summarizing the need for legislation of this nature, Markey recognised:

From housing to health care to national security, algorithms are making consequential decisions, diagnoses, recommendations, and predictions that can significantly alter our lives. As AI supercharges these algorithms, the federal government must protect the marginalized communities that have already been facing the greatest consequences from Big Tech's reckless actions. [The Act] will ensure that the government has the proper tools, resources, and personnel to protect these communities and mitigate AI's dangerous effects.[396]

Current laws focusing on AI with integrity - eliminating bias and social inequalities - all pivot on ensuring that algorithms are designed, developed, deployed, and used ethically. However, probably one of the most significant changes taking root in the emerging legislation is the requirement that developers include bias detection *throughout* the value chain, starting with the first steps of system design. This will be critical to establishing the foundations of responsible technology. That said, one of the biggest challenges faced by lawmakers is to balance regulation that may stifle AI innovation and development, with responsible development

[395] Markey, E. 2023. *Senator Markey* introduces legislation to mandate civil rights offices in federal agencies that handle artificial intelligence. December 12. https://www.markey.senate.gov/news/press-releases/senator-markey-introduces-legislation-to-mandate-civil-rights-offices-in-federal-agencies-that-handles-artificial-intelligence. (Accessed on 22/06/2024).

[396] Markey above fn 76.

and implementation that ensures and assures the fundamental values of equality and human rights.

A second caution for lawmakers to consider is the reporting regulations: "… reliance on companies self-reporting imperils the public or government's ability to catch AI discrimination before it's done harm," state Bedayne et al.[397] Riccardi expresses her own concerns that the well-intended good governance tools, like audits and impact assessments could become "administrative wand-waving". Explaining her disquiet with an example involving a video interview platform used for hiring which was reported to the Federal Trade Commission for its use of algorithms and facial recognition technology, the company hired independent auditors to investigate bias in its product. At the end of the audit, the company issued a statement stating that no bias risks had been found. However, reports Riccardi, when the audit is accessed, "it does not state that the tool is unbiased – just that the auditors didn't have enough information to decide."[398]

A further conundrum is the reality that notwithstanding regulation, full disclosure of how an AI system works is often not possible due to their black box nature. Lawmakers need to specifically factor this into developer and user obligations when dealing with mandatory reporting for transparency and accountability.

12.6 Conclusion

As AI becomes ubiquitous, its potential for bias and discrimination is exponentially increased unless there is responsible management and

[397] Bedayne, J., Haigh, S., Nguyen, T. & Bohrer, B. 2024. *First major attempts to regulate AI face headwinds from all sides.* April 19. 5. https://apnews.com/article/lobbying-artificial-intelligence-regulation-bias-discrimination-58eab95d279648e759f9099ca4249f00. (Accessed on 29/5/2024).

[398] Riccardi above fn 44.

oversight. This cannot be left to chance or even to the developers or implementing industries. The clarion call is for countries to implement legal frameworks responding to the scourge of AI bias and discrimination and promoting transparency, fairness, and accountability. At the heart of the project is *humanity* and the value that is placed on peoples' fundamental rights as human beings. At the AI for Good Summit, July 2023 the UN Secretary-General cemented the sentiment.

[D]eveloping AI for the good of all requires a framework grounded in human rights, transparency, and accountability."[399]

While recognizing the fast-paced development of AI technologies and the generally slower pace of legal innovation, it nevertheless beggars belief that countries committed to the values of human rights would not have in place some AI-specific protections for innocent citizens. Legal provisions excluding a particular AI focus just do not cover the unique challenges posed by AI, and particularly AI-generated biases and discrimination, primarily because of the complexity, enforcement challenges, and inadequate terminology connected with artificial intelligence.[400] Proving discriminatory intent and identifying responsible parties will be two key elements that will continue to impede effective enforcement.

In addition to country-specific laws, the need for global regulation is becoming critical. As pointed out by Zartis: *AI is a global technology. Establishing effective regulations requires international cooperation to*

[399] https://www.un.org/africarenewal/magazine/july-2023/un-chief-says-regulation-needed-ai-%E2%80%98benefit-everyone%E2%80%99. (Accessed on 24/06/2024)

[400] Chiappetta above fn 28, 1.

ensure consistency and prevent potential loopholes for companies operating across different regions with varying legal frameworks.[401]

Concurring, Sindhayach strongly urges a "holistic approach" pointing out that that the absence of international harmonization in AI regulations to address the unique challenges will create loopholes that will be always identified and inevitably exploited.[402]

There is no gainsaying the enormous potential of AI for good and for public and private transformation, but if the development, evolution, and operation are left unchecked, untrammeled, and unguarded, a dystopian future is also possible. AI with integrity requires more effective - and more urgent - legal frameworks, and while there is global acknowledgement of this truth, the legal journeys are still clearly lagging the big AI developers. The domestic and international legal responses, it might be argued, have been (and continue to be) much too reticent to satisfactorily guarantee the required AI system transparency and accountability and warrant citizen protection against the scourges of bias and discrimination from machine learning.

---

[401] Zartis Team. n.d. *AI Legislation in the US: AI Regulation Review.* 6. https://www.zartis.com/ai-regulation-review-ai-regulation-in-the-us/. (Accessed on 29/5/2024).

[402] Sindhayach, R. 2024. Algorithmic bias in AI: Legal landscape and the call for regulations. February 28. *Medium.* https://medium.com/@rupasindhayach/algorithm-bias-in-ai-legal-landscape-and-the-call-for-regulations-47663ff2cf6c. (Accessed on 29/5/2024).

13

AI REGULATION FROM INDIGENOUS MAORI / NEW ZEALAND PERSPECTIVE

Karaitiana Taiuru, New Zealand

13.1 Introduction

This article[403] delves into the intersection of traditional Māori cultural beliefs and practices with the discussion if Artificial Intelligence (AI) regulation is required in New Zealand.

Until only recently, commentators in the AI and regulation space have been bias against Māori perspectives due to no Māori voices being considered, a matter that has now significantly changed with industry representative groups and more Māori in Artificial Intelligence (Taiuru, 2024).

Two primary sets of research are analysed from a 2023 DataCom[404] survey on business leaders use and the recent 2024 InternetNZ survey of New Zealand perceptions of online and AI[405].

---

[403] *Karaitiana Taiuru*, PhD, is a leading authority and a highly accomplished Māori technology ethicist specialising in Māori rights with AI, Māori Data Sovereignty and Governance with emerging digital technologies and biological sciences.

[404] https://datacom.com/nz/en/solutions/experience/insights/ai-attitudes-research-report

[405] https://internetnz.nz/assets/Uploads/New-Zealands-Internet-Insights-2023.pdf

Also considered is research commissioned by The Law Foundation New Zealand and NetSafe[406] research about online abuse that show Māori are more likely to be victims of online abuse in all of its forms.

In conclusion, an analysis of New Zealand legislation and legal frameworks, how they could impact AI regulation in New Zealand.

13.2 New Zealand community perceptions

Recent research by InternetNZ in the report New Zealand's Internet Insights 2023[407] reveals there is a national interest in regulating AI. Māori are more likely than the rest of New Zealand society to want regulation, despite Māori making up 17.3% of the population.

Issue	% New Zealanders	% Māori
Artificial Intelligence (AI's) impact on society	52%	51%
that young children can access inappropriate content	73% o	74%
the security of personal data	69%	68%
Online crime	67%	64%
Identity theft	66%	65%
Information is misleading or wrong	65%	60%

Table 1 New Zealand community perceptions

The report when read as a whole, shows there is an increasing awareness of the risks of technologies such as AI in communities and households. This does provide opportunities for the DigiTech sector to engage

[406] https://netsafe.org.nz/

[407] https://internetnz.nz/assets/Uploads/New-Zealands-Internet-Insights-2023.pdf

more with Māori and communities to seek solutions, correct false statements and to create new roles in AI.

13.3 New Zealand business leaders' opinions of AI regulation

In 2023, DataCom released their report AI Attitudes in New Zealand, AI Attitudes Research, New Zealand. August 2023 DataCom, 2023). The report surveyed 200 senior business leaders in New Zealand to understand rates of AI adoption, maturity of existing AI governance and security, appetite for AI-specific legislation and which sectors are seen to hold the greatest opportunities and risks for AI use.

The need for New Zealand businesses to have the right governance in place to balance the risks and rewards was noted, as was the need to forward-looking approaches from both industry and government to ensure New Zealand remains competitive.

> "AI bias is something organisations need to be particularly mindful of. While AI can be used to support and drive diversity and inclusion initiatives, without the right checks and balances, it has the potential to do the opposite."

50% of New Zealand InfoTech companies surveyed were concerned about bias in AI algorithms that reflect human biases, 73% concerned about AI safety and 53% questioning the ethical use of AI in the wider society.

A staggering 82% of New Zealand InfoTech companies believe that the government should regulate AI while 11% said no, 7% were unsure.

60% of businesses are concerned or do not feel educated about the security of personal data and Intelectual Property with AI. This is a real concern and as I will note later, the current legislation requires an update. Many artists around the world and in New Zealand share these concerns too.

69% of business leaders are supportive of their employees using AI such as ChatGPT (generative AI) to carry out their work tasks in the workplace. A minority of 11% (about 1 in 10) are opposed and 20%, or one in 5 were unsure.

This appears to show that New Zealand businesses are aware of bias and security issues with AI. The result should be that businesses with good governance practices will engage expert stakeholders to regulate against bias and to protect their organisations from litigation and security issues caused by the introduction of AI.

13.4 New Zealand government perspective of AI regulation

Minister for Digitising Government, Science, Innovation and Technology, among other important portfolios, the Hon Judith Collins told the NZ Herald:

> "This Government is committed to getting New Zealand up to speed on AI. We have a cross-party AI caucus, which is due to meet soon. Its first step will be providing feedback on the AI framework we are developing to support responsible and trustworthy AI innovation in government, which the public should expect to hear more on in the coming months. There will be no extra regulation at this stage."[408]

Any sudden reaction without proper considerations of all New Zealand society, business, industry collaborations, innovations, reviews of outdated legislation, and consultation could have negative impacts for

[408] https://www.nzherald.co.nz/business/survey-finds-most-kiwis-worried-about-malicious-ai-technology-minister-judith-collins-responds/DJWKCXXSF5CCH-PPNE47BTCQDJU/

parts of society from commercial, not for profits, minorities, and in particular Māori.

One recent example of this could be seen with the recently repealed Therapeutic Products Act 2023[409]. The repealed Act suggested significant negative impacts on Māori and was likened to the Tohunga Suppression Act 1908 that outlawed many Māori customs including medicines, by many Māori natural medicine practitioners (Norris & Beresford, 2024).

13.5 New Zealand legislation and frameworks

There is a common response that AI does not need to be regulated as New Zealand have current legislation and frameworks to protect New Zealanders. Unfortunately, this is not an opinion that is reflective across all of New Zealand society.

The following rules, regulations and legislation have been commonly touted as suitable for AI usage. Next to each option, I offer an analysis because it is not suitable for Māori.

Privacy Act 1993[410]. This legislation does not recognise traditional Māori cultural perspective and practices with privacy and largely contradicts Māori Data Sovereignty principles (Taiuru, 2022). The Act is solely focused on individual privacy rights with no considerations for collective sharing of data, despite this being recognised in the High Court case Te Pou Matakana Limited v Attorney-General (No 1) [2021] NZHC 3319 (WOCA 2). Nor does it consider the Supreme Cour case Peter Hugh McGregor Ellis v R [2022] NZSC 115, 07 October 2022 where Tikanga

---

409 https://www.legislation.govt.nz/act/public/2023/0037/latest/DLM6914502.html#DLM7312612

410 https://www.legislation.govt.nz/act/public/1993/0028/latest/DLM296639.html

Māori (traditional customary lore) is recognised New Zealand's first common law.

Section 21 (c) of the Act states that "Commissioner to have regard to certain matters take account of cultural perspectives on privacy". At this stage it is too ambiguous, and Māori will need to continue to seek clarification on this section with different circumstances.

The guidelines of the Media Council of New Zealand. These guidelines have been analysed by a number of Māori academics and have consistently found bias to Māori in the media including: How news items represent Māori[411] , Racism and silencing in the media in Aotearoa New Zealand[412] , Our Truth, Tā Mātou Pono: The Press' bias has failed Māori[413], Decolonising the news: 4 fundamental questions media can ask themselves when covering stories about Māori[414]. The guidelines would benefit greatly with a set of Māori perspectives. Media biases can be increased by AI algorithms.

Harmful Digital Communication Act 2015[415]. This Act makes it difficult for any victims to get justice. Māori are more vulnerable than many others and are increasingly becoming a victim group whose cultural values are supressed[416]. Considerations of Artificial Intelligence bullying a human has not been considered, nor if an AI was created to specifically target certain gender, race, cultural, religious groups.

---

411 https://www.iue.net.nz/wp-content/uploads/2023/09/Maori-in-the-Media-Checklist.pdf

412 https://journals.sagepub.com/doi/10.1177/11771801211015436

413 https://www.stuff.co.nz/pou-tiaki/our-truth/300153938/our-truth-t-mtou-pono-the-press-bias-has-failed-mori

414 https://theconversation.com/decolonising-the-news-4-fundamental-questions-media-can-ask-themselves-when-covering-stories-about-maori-205916

415 https://www.legislation.govt.nz/act/public/2015/0063/latest/whole.html

416 https://taiuru.co.nz/category/cyber-safety/

Social Media platform guidelines are just that, guidelines. No Social Media platform guidelines that were sourced for this article consider Te Tiriti o Waitangi or a Māori cultural perspective.

Copyright Act 1994[417]. This Act has never considered Māori intellectual property rights and is already overdue for a review due to emerging technologies. No country in the world has suitably addressed Indigenous collective property rights (WIPO, n.d.).

Further to this, the Copyright Act has been waiting for the findings from the WAI 262 claim which underpins a significant amount of New Zealand's Intellectual Property laws including the Patents Act 2013 and Trademarks Act 2002, recommendations and how to implement them[418]. Of note, we have already seen overseas that AI generated art and the multiple legal issues that arises from this.

There are a number of laws in New Zealand that already require to be updated to adjust to the digital revolution before any new regulation is introduced. It would be an ideal time to also consider Māori and other minorities to guard against bias and discrimination.

13.6 Other considerations

In New Zealand there is a constitution between the indigenous Peoples Māori and the British. The document is called The Treaty of Waitangi Te Tiriit o Waitangi (Te Tiriti) and is legally recognised in legislation (Orange, 2024). Parts of Te Tiriti recognise Māori never ceded sovereignty and that their cultural perspectives must be recognised.

---

417 https://www.legislation.govt.nz/act/public/1994/0143/latest/DLM345634.html

418 https://waitangitribunal.govt.nz/news/ko-aotearoa-tenei-report-on-the-wai-262-claim-released

Based on Te Tirti o Wairangi, 6 T Artificial Intelligence Ethical Principles were developed that must guide any New Zealand regulation (Taiuru, 2023). In addition, the possibility of legal personhood of AI has been discussed using the rights afforded to Māori with Te Tiriti (Taiuru, 2022).

The United Nations Declaration of the Rights of Indigenous Peoples 2007, that New Zealand agreed to in 2010 is also an important legal instrument that recognises Māori rights to technologies such as AI.

Deepfakes are emerging as a new online threat, in particular with elections and pornography that is not considered in current New Zealand legislation. Research commissioned by The Law Foundation New Zealand "Perception Inception: Preparing for deepfakes and the synthetic media of tomorrow" makes a number of recommendations, including

"We recommend caution in developing any substantial new law without first understanding the complex interaction of existing legal regimes. Before acting, it is essential to continue to develop an understanding of how these regimes apply to factual scenarios as they arise. Where new law is necessary, it is likely to take the form of nuanced amendment to existing regulation. For now, existing legislation should be given the opportunity to deal with harms from synthetic media technologies as they arise." (Curtis Barnes, Tom Barraclough, 2019; pg 18).
In addition:

"Any new legislation must take the position that synthetic media technologies and artefacts touch upon individual rights of privacy and freedom of expression, deserving careful attention from policymakers and broad public consultation. There are benefits, risks, and trade-offs to be discussed in deciding whether to allocate responsibility for restricting synthetic media technologies to the State or to private actors. Human rights, the rule of law, natural justice, transparency and accountability are

essential ingredients in whatever approach is adopted." (Curtis Barnes, Tom Barraclough, 2019; pg 18).

13.7 New Zealand algorithm charter

The only AI-specific policy New Zealand has, is the Algorithm Charter, which many Government agencies have blindly signed up to. There have been many criticisms that reflect the lack of consultation with Māori. The Charter doesn't really engage with Māori, including in terms of Māori data sovereignty (Chen, 2022; 2022b); Taiuru, 2020). Taylor Fry reviewed that Algorithm Charter for Aotearoa New Zealand after its first year in December 2021and made a number of concerning findings, but largely noting it did not consider Te Tirti in its findings, only as a reference.

Investigative journalist Phil Pennington reported that compliance was very light-handed.

"Some greater enforcement might be necessary to keep social licence." This appears at odds with OECD principles on artificial intelligence that say how systems work should be disclosed so people can challenge them (Pennington, 2022).

Many of the signatories continue to be accused of bias and racism against Māori. Noting government algorithms impact the individual as well as collectives, the Algorithm Charter could benefit from a wider consultation with Māori societal groups, AI industry groups community groups. This would ensure that wide and reflective Māori society perspectives are captured. In partnerships, the Charter could be an effective way of working with expert stakeholders, industry and government and would avoid the need to consider regulation.

13.8 Conclusion

Currently New Zealand has no regulation of AI. While businesses want regulation, they have yet to implement business policies within their own organisations.

Academic researchers and Māori community perspectives are that regulation of AI is not currently needed in New Zealand. What is required is for the AI industry, legal experts and expert stakeholders such as Māori and community organisations to work together to identify existing laws and amend them and to include a societal perspective that also included Te Tiriti. It is also an opportunity to consider national data sovereignty and Māori Data sovereignty principles (Taiuru, 2022).

Any sudden regulation of AI in New Zealand would likely be premature and cause a number of unintended consequences to Māori and other areas of New Zealand society. If at some future date regulation is considered, then it must be in a consultative and fact-based research justifying why regulation of AI is required and who it will impact.

References

Chen, A. (2022, September 18). The Algorithm Charter. Perspective. University of Auckland. Retrieved from https://informed-futures.org/the-algorithm-charter/

Chen, A. (2022b). The Algorithm Charter | He Tutohi Hatepe mo Aotearoa. In Pendergrast, A. & Pendergrast, K. (2022). (Eds) of More Zeros and Ones Digital Technology, Maintenance and Equity in Aotearoa New Zealand. Bridget Williams

Claudia Orange, 'Te Tiriti o Waitangi – the Treaty of Waitangi', Te Ara - the Encyclopedia of New Zealand, http://www.TeAra.govt.nz/en/te-tiriti-o-waitangi-the-treaty-of-waitangi/print (accessed 18 June 2024)

Curtis Barnes, Tom Barraclough. (2019). PERCEPTION INCEPTION Preparing for deepfakes and the synthetic media of tomorrow.

Fry Taylor. (2021) Algorithm Charter for Aotearoa New Zealand Year 1 Review. Retrieved from https://www.data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-Year-1-Review-FINAL.pdf

Pauline Norris and Rosemary Beresford, 'Medicines and remedies - Plant extracts to modern drugs, 1900 to 1930s', Te Ara - the Encyclopedia of New Zealand, http://www.TeAra.govt.nz/en/document/28223/tohunga-suppression-act (accessed 18 June 2024)

Phil Pennington (2022, 18 September). Government's algorithm charter 'okay, could do better' - review. RNZ. Retrieved https://www.rnz.co.nz/news/national/475011/government-s-algorithm-charter-okay-could-do-better-review

Taiuru, K. (2020, August 22). NZ Algorithm charter – Are Māori protected? Te Kete o Karaitiana Taiuru (Blog)

Taiuru, K. (2022). A Māori Cultural Perspective of AI/Machine Sentience. In Goffi E. R., Momcilovic A., et al. (Eds). Can an AI be sentient? Multiple perspectives on sentience and on the potential ethical implications of the rise of sentient AI. Notes n° 2, (2022). Global AI Ethics Institute.

Taiuru, K. (2022, February 12). Compendium of Māori Data Sovereignty Version 2. Te Kete o Karaitiana Taiuru (Blog)

Taiuru, K. (2023, October 27). 6 Te Tiriti Based Artificial Intelligence Ethical Principles. Te Kete o Karaitiana Taiuru (Blog)

Taiuru, K. (2024, March 24). Māori Voices in the Artificial Intelligence (AI) Landscape of Aotearoa New Zealand. Te Kete o Karaitiana Taiuru (Blog)

WIPO. N.d. Leaflet No. 12: “WIPO and Indigenous Peoples”, Geneva: Global Intellectual Property Issues Division of World Intellectual Property Organization (WIPO) indileafl et12.doc at 1. Available

online at https://www.ohchr.org/Documents/Publications/GuideIPleaflet12en.pdf

PART III

PUBLIC SECTOR AI GUIDELINES, STANDARDS, REGULATIONS AND ETHICS: GLOBAL, CONTINTENTAL, NATIONAL

The public sector with international (UN) agencies, continental Unions and individual governments, mainly on national level, approved in the last few years manifold conventions, guidelines, principles up to legally binding legislations. Private sector companies decided in parallel to the public sector their own internal benchmarks, principles and guidelines (below Part IV). This Part III offers – pars pro toto – a representative selection of key documents: the United Nations and their specialised organisations such as Unesco, continental unions such as European Union, African Union and the Council of Europe as well as the two Nations China and the USA who dominate the AI development.

14

UNITED NATIONS: GOVERNING AI FOR HUMANITY

14.1 Introduction

The United Nations Secretary-General convened a multi-stakeholder High-level Advisory Body on AI on 26 October 2023 to undertake analysis and advance recommendations for the international governance of AI by a globally inclusive approach. The AI Advisory Body comprised 39 preeminent AI leaders from 33 countries from across all regions and multiple sectors, serving in their personal capacity. The Advisory Body presented in a record speed of only two months, a draft report in December 2023 with inputs from 2000 AI experts, 18 deep-dive discussions, more than 50 consultation sessions across all regions, more than 250 written submissions from over 150 organizations and 100 individuals. The AI Advisory Body presented its Final Report "Governing AI for Humanity"[419] in September 2024. The following text is the Executive Summary of the report (pp.7-21).

The Editors

---

[419] https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

14.2 Governing AI for humanity. Executive summary

(i) Artificial intelligence (AI) is transforming our world. This suite of technologies offers tremendous potential for good, from opening new areas of scientific inquiry and optimizing energy grids, to improving public health and agriculture and promoting broader progress on the Sustainable Development Goals (SDGs). **(ii)** Left ungoverned, however, AI's opportunities may not manifest or be distributed equitably. Widening digital divides could limit the benefits of AI to a handful of States, companies and individuals. Missed uses – failing to take advantage of and share AI-related benefits because of lack of trust or missing enablers such as capacity gaps and ineffective governance – could limit the opportunity envelope. **(iii)** AI also brings other risks. AI bias and surveillance are joined by newer concerns, such as the confabulations (or "hallucinations") of large language models, AI-enhanced creation and dissemination of disinformation, risks to peace and security, and the energy consumption of AI systems at a time of climate crisis. **(iv)** Fast, opaque and autonomous AI systems challenge traditional regulatory systems, while ever-more-powerful systems could upend the world of work. Autonomous weapons and public security uses of AI raise serious legal, security and humanitarian questions. **(v)** There is, today, a global governance deficit with respect to AI. Despite much discussion of ethics and principles, the patchwork of norms and institutions is still nascent and full of gaps. Accountability is often notable for its absence, including for deploying non-explainable AI systems that impact others. Compliance often rests on voluntarism; practice belies rhetoric. 1 See https://un.org/ai-advisory-body. **(vi)** As noted in our interim report,1 AI governance is crucial – not merely to address the challenges and risks, but also to ensure that we harness AI's potential in ways that leave no one behind.

14.2.1 The need for global governance

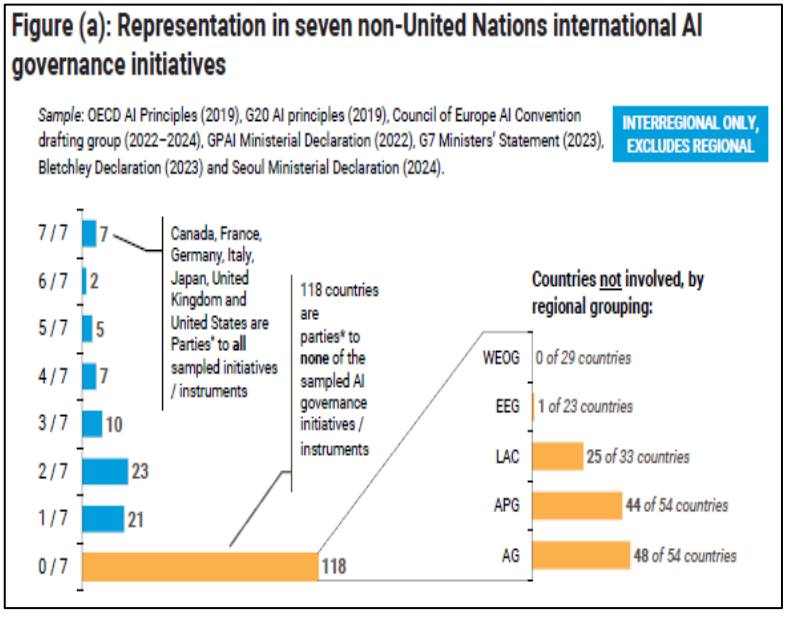
(vii) The imperative of global governance, in particular, is irrefutable. AI's raw materials, from critical minerals to training data, are globally sourced. General-purpose AI, deployed across borders, spawns manifold applications globally. The accelerating development of AI concentrates power and wealth on a global scale, with geopolitical and geoeconomic implications. **(viii)** Moreover, no one currently understands all of AI's inner workings enough to fully control its outputs or predict its evolution. Nor are decision makers held accountable for developing, deploying or using systems they do not understand. Meanwhile, negative spillovers and downstream impacts resulting from such decisions are also likely to be global. **(ix)** The development, deployment and use of such a technology cannot be left to the whims of markets alone. National governments and regional organizations will be crucial, but the very nature of the technology itself – transboundary in structure and application – necessitates a global approach. Governance can also be a key enabler for AI innovation for the SDGs globally. **(x)** AI, therefore, presents challenges and opportunities that require a holistic, global approach cutting transversally across political, economic, social, ethical, human rights, technical, environmental and other domains. Such an approach can turn a patchwork of evolving initiatives into a coherent, interoperable whole, grounded in international law and the SDGs, adaptable across contexts and over time. **(xi)** In our interim report, we outlined principles2 that should guide the formation of new international AI governance institutions. These principles acknowledge that AI governance does not take place in a vacuum, that international law, especially international human rights law, applies in relation to AI.

14.2.2 Global AI governance gaps

(xii) There is no shortage of documents and dialogues focused on AI governance. Hundreds of guides, frameworks and principles[420] have been adopted by governments, companies and consortiums, and regional and international organizations. **(xiii)** Yet, none of them can be truly global in reach and comprehensive in coverage. This leads to problems of representation, coordination and implementation. **(xiv)** In terms of representation, whole parts of the world have been left out of international AI governance conversations. Figure (a) shows seven prominent, non-United Nations AI initiatives. Seven countries are parties to all the sampled AI governance efforts, whereas 118 countries are parties to none (primarily in the global South). **(xv)** Equity demands that more voices play meaningful roles in decisions about how to govern technology that affects us. The concentration of decision-making in the AI technology sector cannot be justified; we must also recognize that historically many communities have been entirely excluded from AI governance conversations that impact them. **(xvi)** AI governance regimes must also span the globe to be effective — effective in averting "AI arms races" or a "race to the bottom" on safety and rights, in detecting and responding to incidents emanating from decisions along AI's life cycle which span multiple jurisdictions, in spurring learning, in encouraging interoperability, and in sharing AI's benefits. The technology is borderless and, as it spreads, the illusion that any one State or group of States could (or should) control it

[420] *Guiding principle* 1: AI should be governed inclusively, by and for the benefit of all; guiding principle 2: AI must be governed in the public interest; guiding principle 3: AI governance should be built in step with data governance and the promotion of data commons; guiding principle 4: AI governance must be universal, networked and rooted in adaptive multi-stakeholder collaboration; guiding principle 5: AI governance should be anchored in the Charter of the United Nations, international human rights law and other agreed international commitments, such as the SDGs.

will diminish. **(xvii)** Coordination gaps between initiatives and institutions risk splitting the world into disconnected and incompatible AI governance regimes. Coordination is also lacking within the United Nations system. Although many United Nations entities touch on AI governance, their specific mandates mean that none does so in a comprehensive manner. **(xviii)** However, representation and coordination are not enough. Accountability requires implementation so that commitments to global AI governance translate to tangible outcomes in practice, including on capacity development and support to small and medium enterprises, so that opportunities are shared. Much of this will take place at the national and regional levels, but more is also needed globally to address risks and harness benefits.



14.2.3 Enhancing global cooperation

(xix) Our recommendations advance a holistic vision for a globally networked, agile and flexible approach to governing AI for humanity,

encompassing common understanding, common ground and common benefits. Only such an inclusive and comprehensive approach to AI governance can address the multifaceted and evolving challenges and opportunities AI presents on a global scale, promoting international stability and equitable development. **(xx)** Guided by principles established in our interim report, our proposals seek to fill gaps and bring coherence to the fast-emerging ecosystem of international AI governance responses and initiatives, helping to avoid fragmentation and missed opportunities. To support these measures efficiently and to partner effectively with other institutions, we propose a light, agile structure as an expression of coherent effort: an AI office in the United Nations Secretariat, close to the Secretary General, working as the "glue" to unite the initiatives proposed here efficiently and sustainably.

14.2.3.1 A) Common understanding

(xxi) A global approach to governing AI starts with a common understanding of its capabilities, opportunities, risks and uncertainties. There is a need for timely, impartial and reliable scientific knowledge and information about AI so that Member States can build a shared foundational understanding worldwide, and to balance information asymmetries between companies housing expensive AI labs and the rest of the world (including via information-sharing between AI companies and the broader AI community). **(xxii)** Pooling scientific knowledge is most efficient at the global level, enabling joint investment in a global public good, and public interest collaboration across otherwise fragmented and duplicative

(xxiii) Learning from precedents such as the Intergovernmental Panel on Climate Change (IPCC) and the United Nations Scientific Committee on the Effects of Atomic Radiation, an international, multidisciplinary scientific panel on AI could collate and catalyse leading-edge research to inform scientists, policymakers, Member States and other stakeholders

seeking scientific perspectives on AI technology or its applications from an impartial, credible source. **xxiv** A scientific panel under the auspices of the United Nations could source expertise on AI-related opportunities. This might include facilitating "deep dives" into applied domains of the SDGs, such as health care, energy, education, finance, agriculture, climate, trade and employment. **xxv** Risk assessments could also draw on the work of other AI research initiatives, with the United Nations offering a uniquely trusted "safe harbour" for researchers to exchange ideas on the "state of the art". By pooling knowledge across silos in countries or companies that may not otherwise engage or be included, a United Nations-hosted panel can help to rectify misperceptions and bolster trust globally. **xxvi** Such a panel should operate independently, with support from a cross-United Nations system team drawn from the below-proposed AI office and relevant United Nations agencies, such as the International Telecommunication Union (ITU) and the United Nations Educational, Scientific and Cultural Organization (UNESCO). It should partner with research efforts led by other international institutions, such as the Organisation for Economic Development OECD and the Global Partnership for Artificial Intelligence.

Recommendation 1: **An international scientific panel on AI.**

We recommend the creation of an independent international scientific panel on AI, made up of diverse multidisciplinary experts in the field serving in their personal capacity on a voluntary basis. Supported by the proposed United Nations AI office and other relevant United Nations agencies, partnering with other relevant international organizations, its mandate would include: **a)** Issuing an annual report surveying AI-related capabilities, opportunities, risks and uncertainties, identifying areas of scientific consensus on technology trends and areas where additional research is needed; **b)** Producing quarterly thematic research digests on areas in which AI could help to achieve the SDGs, focusing on areas of

public interest which may be under-served; and **c)** Issuing ad hoc reports on emerging issues, in particular the emergence of new risks or significant gaps in the governance landscape.

14.2.3.2 B) Common ground

xxvii Alongside a common understanding of AI, common ground is needed to establish interoperable governance approaches anchored in global norms and principles in the interests of all countries. This is required at the global level to avert regulatory races to the bottom while reducing regulatory friction across borders; to maximize learning and technical interoperability; and to respond effectively to challenges arising from the transboundary character of AI.

Policy dialogue on AI governance: xxviii An inclusive policy forum is needed so that all Member States, drawing on the expertise of stakeholders, can share best practices that are based on human rights and foster development, that foster interoperable governance approaches and that account for transboundary challenges that warrant further policy consideration. This does not mean global governance of all aspects of AI, but it can set the framework for international cooperation and better align industry and national efforts with global norms and principles. **xxix** Institutionalizing such multi-stakeholder exchange under the auspices of the United Nations can provide a reliably inclusive home for discussing emerging governance practices and appropriate policy responses. By edging beyond comfort zones, dialogue between non-likeminded countries, and between States and stakeholders, can catalyse learning and lay foundations for greater cooperation, such as on safety standards and rights, and for times of global crisis. A United Nations setting is essential to anchoring this effort in the widest possible set of shared norms. **xxx** Combined with capacity development (see recommendations 4 and 5), such inclusive dialogue on governance approaches can help States and companies to update their regulatory approaches and methodologies to

respond to accelerating AI. Connections to the international scientific panel would enhance that dynamic, comparable to the relationship between IPCC and the United Nations Climate Change Conference. **xxxi** A policy dialogue could begin on the margins of existing meetings in New York (such as the General Assembly4) and in Geneva. Twice-yearly meetings could focus more on opportunities across diverse sectors in one meeting, and more on risks in the other meeting. Moving forward, a gathering like this would be an appropriate forum for sharing information about AI incidents, such as those that stretch or exceed the capacities of existing agencies. **xxxii** One portion of each dialogue session might focus on national approaches led by Member States, with a second portion sourcing expertise and inputs from key stakeholders – in particular, technology companies and civil society representatives. In addition to the formal dialogue sessions, multi-stakeholder engagement on AI policy could leverage other existing, more specialized mechanisms, such as the ITU AI for Good meeting, the annual Internet Governance Forum meeting, the UNESCO Global Forum on AI Ethics and the United Nations Conference on Trade and Development (UNCTAD) e-Week.

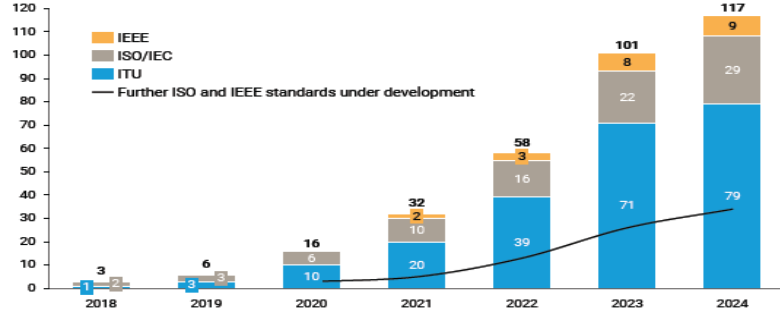
Recommendation 2: **Policy dialogue on AI governance**

We recommend the launch of a twice-yearly intergovernmental and multi-stakeholder policy dialogue on AI governance on the margins of existing meetings at the United Nations. Its purpose would be to: **a)** Share best practices on AI governance that foster development while furthering respect, protection and fulfilment of all human rights, including pursuing opportunities as well as managing risks; **b)** Promote common understandings on the implementation of AI governance measures by private and public sector developers and users to enhance international interoperability of AI governance; **c)** Share voluntarily significant AI incidents that stretched or exceeded the capacity of State agencies to respond; and

d) Discuss reports of the international scientific panel on AI as appropriate.

AI standards exchange: **xxxiii** When AI systems were first explored, few standards existed to help to navigate or measure this new frontier. More recently, there has been a proliferation of standards. Figure (b) illustrates the increasing number of standards adopted by ITU, the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC) and the Institute of Electrical and Electronics Engineers (IEEE). **xxxiv** There is no common language among these standards bodies, and many terms routinely used with respect to AI – fairness, safety, transparency – do not have agreed definitions. There are also disconnects between those standards that were adopted for narrow technical or internal validation purposes, and those that are intended to incorporate broader ethical principles. We now have an emerging set of standards that are not grounded in a common understanding of meaning or are divorced from the values that they were intended to uphold. **xxxv** Drawing on the expertise of the international scientific panel and incorporating members from the various national and international entities that have contributed to standard-setting, as well as representatives from technology companies and civil society, the United Nations system could serve as a clearing house for AI standards that would apply globally.

Figure (b): Number of standards related to AI



Recommendation 3: AI standards exchange.

We recommend the creation of an AI standards exchange, bringing together representatives from national and international standard-development organizations, technology companies, civil society and representatives from the international scientific panel. It would be tasked with: **a)** Developing and maintaining a register of definitions and applicable standards for measuring and evaluating AI systems; **b)** Debating and evaluating the standards and the processes for creating them; and **c)** Identifying gaps where new standards are needed.

14.2.3.3 C) Common benefits

xxxvi The 2030 Agenda for Sustainable Development, with its 17 SDGs, can give clarity of purpose to the development, deployment and uses of AI, bending the arc of investments towards global development challenges. Without a comprehensive and inclusive approach to AI governance, the potential of AI to contribute positively to the SDGs could be missed, and its deployment could inadvertently reinforce or exacerbate disparities and biases. **xxxvii** AI is no panacea for sustainable development challenges; it is one component within a broader set of solutions. To truly unlock AI's potential to address societal challenges, collaboration among governments, academia, industry and civil society is crucial, so that AI-enabled solutions are inclusive and equitable. **xxxviii** Much of this depends on access to talent, computational power (or "compute") and data, in ways that help cultural and linguistic diversity to flourish. Basic infrastructure and the resources to maintain it are also pre-requisites. **xxxix** Regarding talent, not every society needs cadres of computer scientists for building their own models. However, whether technology is bought, borrowed or built, a baseline socio-technical capacity is needed to understand the capabilities and limitations of AI, and harness AI-enabled use cases appropriately while addressing context-specific risks. **xl** Compute is one of the biggest barriers to entry in the field of AI.

Of the top 100 high-performance computing clusters in the world capable of training large AI models, not one is hosted in a developing country. It is unrealistic to promise access to compute that even the wealthiest countries and companies struggle to acquire. Rather, we seek to put a floor under the AI divide for those unable to secure needed enablers via other means, including by supporting initiatives towards distributed and federated AI development models. **xli** Turning to data, it is common to speak of misuse of data in the context of AI (such as infringements on privacy) or missed uses of data (failing to exploit existing data sets). However, a related problem is missing data, which includes the large portions of the globe that are data poor. Failure to reflect the world's linguistic and cultural diversity has been linked to bias in AI systems, but may also be a missed opportunity for those communities to access AI's benefits. **xlii** A set of shared resources – including open models – is needed to support inclusive and effective participation by all Member States in the AI ecosystem, and here global approaches have distinct advantages.

Capacity development network: xliii Growing public and private demand for human and other AI capacity coincides with emergent national, regional and public-private AI centres of excellence that have international capacity development roles. A global network can serve as a matching platform that expands the range of possible partnering and enhances interoperability of capacity-building approaches. **xliv** From the Millennium Development Goals to the SDGs, the United Nations has long embraced developing the capacities of individuals and institutions. A network of institutions, affiliated with the United Nations, could expand options for countries seeking capacity partnerships. It could also catalyse new national centres of excellence to stimulate the development of local AI innovation ecosystems, following interoperable approaches aligned with United Nations normative commitments. **xlv** Such a network would promote an alternative paradigm of AI technology development:

bottom-up, cross-domain, open and collaborative. National-level efforts could continue to employ diagnosis tools, such as the UNESCO AI Readiness Assessment Methodology, to help to identify gaps at the national level, with the international network helping to address them.

Recommendation 4: **Capacity development network**

We recommend the creation of an AI capacity development network to link up a set of collaborating, United Nations-affiliated capacity development centres making available expertise, compute and AI training data to key actors. The purpose of the network would be to: **a)** Catalyse and align regional and global AI capacity efforts by supporting networking among them; **b)** Build AI governance capacity of public officials to foster development while furthering respect, protection and fulfilment of all human rights; **c)** Make available trainers, compute and AI training data across multiple centres to researchers and social entrepreneurs seeking to apply AI to local public interest use cases, including via: **i)** Protocols to allow cross-disciplinary research teams and entrepreneurs in compute-scarce settings to access compute made available for training/tuning and applying their models appropriately to local contexts; **ii)** Sandboxes to test potential AI solutions and learn by doing; **iii)** A suite of online educational opportunities on AI targeted at university students, young researchers, social entrepreneurs and public sector officials; and **iv)** A fellowship programme for promising individuals to spend time in academic institutions or technology companies.

Global fund for AI: xlvi Many countries face fiscal and resource constraints limiting their ability to use AI appropriately and effectively. Despite any capacity development efforts (recommendation 4), some may still be unable to access training, compute, models and training data without international support. Other funding efforts may also not scale without it. **xlvii** Our intention in proposing a fund is not to guarantee access to advanced compute resources and capabilities. The answer may

not always be more compute. We also need better ways to connect talent, compute and data. The fund's purpose would be to address the underlying capacity and collaboration gaps for those unable to access requisite enablers so that: a. Countries in need can access AI enablers, putting a floor under the AI divide; b. Collaborating on AI capacity development leads to habits of cooperation and mitigates geopolitical competition; c. Countries with divergent regulatory approaches have incentives to develop common templates for governing data, models and applications for societal-level challenges related to the SDGs and scientific breakthroughs. **xlviii** This public interest focus makes the fund complementary to the proposal for an AI capacity development network, to which the fund would also channel resources. The fund would provide an independent capacity for monitoring of impact, and could source and pool in-kind contributions, including from private sector entities, to make available AI-related training programmes, time, compute, models and curated data sets at lower-thanmarket cost. In this manner, we ensure that vast swathes of the world are not left behind and are instead empowered to harness AI for the SDGs in different contexts. **xlix** It is in everyone's interest to ensure that there is cooperation in the digital world as in the physical world. Analogies can be made to efforts to combat climate change, where the costs of transition, mitigation or adaptation do not fall evenly, and international assistance is essential to help resource-constrained countries so that they can join the global effort to tackle a planetary challenge.

Recommendation 5: **Global fund for AI**

We recommend the creation of a global fund for AI to put a floor under the AI divide. Managed by an independent governance structure, the fund would receive financial and in-kind contributions from public and private sources and disburse them, including via the capacity development network, to facilitate access to AI enablers to catalyse local empowerment for the SDGs, including: **a)** Shared computing resources for

model training and fine-tuning by AI developers from countries without adequate local capacity or the means to procure it; **b)** Sandboxes and benchmarking and testing tools to mainstream best practices in safe and trustworthy model development and data governance; **c)** Governance, safety and interoperability solutions with global applicability; **d)** Data sets and research into how data and models could be combined for SDG-related projects; and **e)** A repository of AI models and curated data sets for the SDGs.

Global AI data framework

l Access to AI training data, via market or other mechanisms, is a critical enabler for flourishing local AI ecosystems — particularly in countries, communities, regions and demographic groups with "missing" data (see the section on "common benefits" above). **li** Only global collective action can incentivize interoperability, stewardship, privacy preservation, empowerment and rights enhancement in ways that promote a "race to the top" across jurisdictions towards protection of human rights and other agreed commitments, data availability and fair compensation to data subjects in the governance of the collection, creation, use and monetization of AI training data. This aim motivates our proposal for a global AI data framework. **lii** Such a framework would not create new data-related rights. Rather, it would address issues of availability, interoperability and use of AI training data. It would help to build common understanding on how to align different national and regional data protection frameworks. It could also promote flourishing local AI ecosystems supporting cultural and linguistic diversity, as well as limiting further economic concentration. **liii** These measures could be complemented by promoting data commons and provisions for hosting data trusts in areas relevant to the SDGs, based on templates for agreements to hold and **18** Governing AI for Humanity share data in a fair, safe and equitable manner. The development of these templates and the actual storage and

analysis of data held in commons or in trusts could be supported by the proposed capacity development network and global fund for AI (recommendations 4 and 5). **liv** The United Nations is uniquely positioned to support the establishment of global principles and practical arrangements for AI training data governance and use, in line with agreed international commitments on human rights, intellectual property and sustainable development, building on years of work by the data community and integrating it with recent developments on AI ethics and governance. This is analogous to the role of the United Nations Commission on International Trade Law in advancing international trade by developing legal and non-legal cross-border frameworks. **lv** Similarly, the Commission on Science and Technology for Development and the Statistical Commission have on their agenda data for development and data on the SDGs. There are also important issues of content, copyright and protection of indigenous knowledge and cultural expression being considered by the World Intellectual Property Organization (WIPO).

Recommendation 6: **Global AI data framework**

We recommend the creation of a global AI data framework, developed through a process initiated by a relevant agency such as the United Nations Commission on International Trade Law and informed by the work of other international organizations, for: **a)** Outlining data-related definitions and principles for global governance of AI training data, including as distilled from existing best practices, and to promote cultural and linguistic diversity; **b)** Establishing common standards around AI training data provenance and use for transparent and rights-based accountability across jurisdictions; and **c)** Instituting market-shaping data stewardship and exchange mechanisms for enabling flourishing local AI ecosystems globally, such as: **i)** Data trusts; **ii)** Well-governed global marketplaces for exchange of anonymized data for training AI models; and **iii)** Model agreements for facilitating international data access and

global interoperability, potentially as techno-legal protocols to the framework.

14.2.3.4 D) Coherent effort

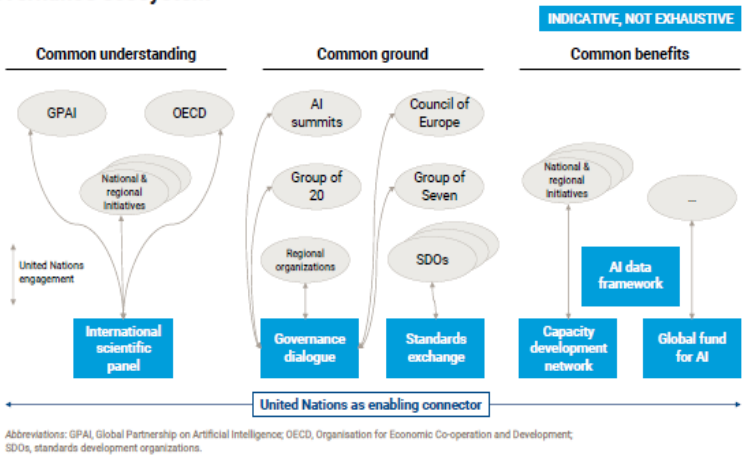
lvi The above proposals seek to address the representation, coordination and implementation gaps identified in the emerging international AI governance regime. These gaps can be addressed through partnerships and collaboration with existing institutions and mechanisms to promote a common understanding, common ground and common benefits. **lvii** Nevertheless, without a dedicated focal point in the United Nations to support and enable coordination among these and other efforts, the world will lack the inclusively networked, agile and coherent approach required for effective and equitable governance of AI as a transboundary, fast-changing and general-purpose technology. **lviii** The patchwork of norms and institutions outlined under the section "Global AI governance gaps" above, reflect widespread recognition that governance of AI is a global necessity. The unevenness of that response demands some measure of coherent effort.

AI office in the United Nations Secretariat

lix We, therefore, propose a light touch mechanism to act as the "glue" that supports and catalyses the proposals in this report, including through partnerships, while also enabling the United Nations system to speak with one voice in the evolving AI governance ecosystem. **lx** This small, agile capacity, in the form of an AI office within the United Nations Secretariat, would report to the Secretary-General, conferring the benefit of connections throughout the United Nations system, without being tied to one part of it. That is important because of the uncertain future of AI and the strong likelihood that it will permeate all aspects of human endeavour. **lxi** Such a body should be agile, champion inclusion and partner rapidly to accelerate coordination and implementation – drawing as a

first priority on existing resources and functions within the United Nations system. The focus should be on civilian applications of AI.

Figure (c): Proposed role of the United Nations in the international AI governance ecosystem



Abbreviations: GPAI, Global Partnership on Artificial Intelligence; OECD, Organisation for Economic Co-operation and Development; SDOs, standards development organizations.

lxiii It could be staffed in part by United Nations personnel seconded from specialized agencies and other parts of the United Nations system, such as ITU, UNESCO, the Office of the United Nations High Commissioner for Human Rights (OHCHR), UNCTAD, the United Nations University and the United Nations Development Programme (UNDP). It should engage multiple stakeholders, including companies, civil society and academia, and work in partnership with leading organizations outside of the United Nations (see fig. (c)). This would position the United Nations to enable connections for fostering common understanding, common ground and common benefits in the international AI governance ecosystem. **lxiii** Recommendation 7 is made on the basis of a clear-eyed assessment as to where the United Nations can add value, including where it can lead, where it can aid coordination and where it should step aside. It also brings the benefits of existing institutional

arrangements, including pre-negotiated funding and administrative processes that are well established and understood.

Recommendation 7: **AI office within the Secretariat**

We recommend the creation of an AI office within the Secretariat, reporting to the Secretary General. It should be light and agile in organization, drawing, wherever possible, on relevant existing United Nations entities. Acting as the "glue" that supports and catalyses the proposals in this report, partnering and interfacing with other processes and institutions, the office's mandate would include: **a)** Providing support for the proposed international scientific panel, policy dialogue, standards exchange, capacity development network and, to the extent required, the global fund and global AI data framework; **b)** Engaging in outreach to diverse stakeholders, including technology companies, civil society and academia, on emerging AI issues; and **c)** Advising the Secretary-General on matters related to AI, coordinating with other relevant parts of the United Nations system to offer a whole-of United Nations response.

15

UNESCO AI VALUES AND PRINCIPLES

15.1 Introduction

Unesco, the leading UN agency for education and science, works since a long time on the Ethics of Artificial Intelligence and on governance rules. The following text is an excerpt of the important Unesco document "Recommendation on the Ethics of Artificial Intelligence", published by Unesco 2022. The following excerpt is the full chapter III "Values and Principles"[421.] The full Unesco text[422] includes 1. Scope of Application, 2. Aims and Objectives, 3. Values and Principles, 4. Areas of Policy Action, 5. Monitoring and Evaluation, 6. Utilization and Exploitation of the Present Recommendation, 7. Promotion of the Present Recommendation, 8. Final Provision.

The Editors

15.2 Unesco AI values and principles

9. The values and principles included below should be respected by all actors in the AI system life cycle, in the first place and, where needed and appropriate, be promoted through amendments to the existing and

---

[421] Unesco, *Recommendation on the Ethics of Artificial Intelligence,* adopted on 23 Nov. 2021, published by Unesco, Paris 2022*,* chapter III, 17-23.

[422] https://unesdoc.unesco.org/ark:/48223/pf0000381137.

elaboration of new legislation, regulations and business guidelines. This must comply with international law, including the United Nations Charter and Member States' human rights obligations, and should be in line with internationally agreed social, political, environmental, educational, scientific and economic sustainability objectives, such as the United Nations Sustainable Development Goals (SDGs).

10. Values play a powerful role as motivating ideals in shaping policy measures and legal norms. While the set of values outlined below thus inspires desirable behaviour and represents the foundations of principles, the principles unpack the values underlying them more concretely so that the values can be more easily operationalized in policy statements and actions.

11. While all the values and principles outlined below are desirable per se, in any practical contexts, there may be tensions between these values and principles. In any given situation, a contextual assessment will be necessary to manage potential tensions, taking into account the principle of proportionality and in compliance with human rights and fundamental freedoms. In all cases, any possible limitations on human rights and fundamental freedoms must have a lawful basis, and be reasonable, necessary and proportionate, and consistent with States' obligations under international law. To navigate such scenarios judiciously will typically require engagement with a broad range of appropriate stakeholders, making use of social dialogue, as well as ethical deliberation, due diligence and impact assessment.

12. The trustworthiness and integrity of the life cycle of AI systems is essential to ensure that AI technologies will work for the good of humanity, individuals, societies and the environment and ecosystems, and embody the values and principles set out in this Recommendation. People should have good reason to trust that AI systems can bring individual and shared benefits, while adequate measures are taken to mitigate risks.

An essential requirement for trustworthiness is that, throughout their life cycle, AI systems are subject to thorough monitoring by the relevant stakeholders as appropriate. As trustworthiness is an outcome of the operationalization of the principles in this document, the policy actions proposed in this Recommendation are all directed at promoting trustworthiness in all stages of the AI system life cycle.

15.3 Unesco AI values

(Chapter III.1) **Respect, protection and promotion of human rights and fundamental freedoms and human dignity**

13. The inviolable and inherent dignity of every human constitutes the foundation for the universal, indivisible, inalienable, interdependent and interrelated system of human rights and fundamental freedoms. Therefore, respect, protection and promotion of human dignity and rights as established by international law, including international human rights law, is essential throughout the life cycle of AI systems. Human dignity relates to the recognition of the intrinsic and equal worth of each individual human being, regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.

14. No human being or human community should be harmed or subordinated, whether physically, economically, socially, politically, culturally or mentally during any phase of the life cycle of AI systems. Throughout the life cycle of AI systems, the quality of life of human beings should be enhanced, while the definition of "quality of life" should be left open to individuals or groups, as long as there is no violation or abuse of human rights and fundamental freedoms, or the dignity of humans in terms of this definition.

15. Persons may interact with AI systems throughout their life cycle and receive assistance from them, such as care for vulnerable people or people in vulnerable situations, including but not limited to children, older persons, persons with disabilities or the ill. Within such interactions, persons should never be objectified, nor should their dignity be otherwise undermined, or human rights and fundamental freedoms violated or abused.

16. Human rights and fundamental freedoms must be respected, protected and promoted throughout the life cycle of AI systems. Governments, private sector, civil society, international organizations, technical communities and academia must respect human rights instruments and frameworks in their interventions in the processes surrounding the life cycle of AI systems. New technologies need to provide new means to advocate, defend and exercise human rights and not to infringe them.

Environment and ecosystem flourishing

17. Environmental and ecosystem flourishing should be recognized, protected and promoted through the life cycle of AI systems. Furthermore, environment and ecosystems are the existential necessity for humanity and other living beings to be able to enjoy the benefits of advances in AI.

18. All actors involved in the life cycle of AI systems must comply with applicable international law and domestic legislation, standards and practices, such as precaution, designed for environmental and ecosystem protection and restoration, and sustainable development. They should reduce the environmental impact of AI systems, including but not limited to its carbon footprint, to ensure the minimization of climate change and environmental risk factors, and prevent the unsustainable exploitation, use and transformation of natural resources contributing to the deterioration of the environment and the degradation of ecosystems.

Ensuring diversity and inclusiveness

19. Respect, protection and promotion of diversity and inclusiveness should be ensured throughout the life cycle of AI systems, consistent with international law, including human rights law. This may be done by promoting active participation of all individuals or groups regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin.

20. The scope of lifestyle choices, beliefs, opinions, expressions or personal experiences, including the optional use of AI systems and the co-design of these architectures should not be restricted during any phase of the life cycle of AI systems.

21. Furthermore, efforts, including international cooperation, should be made to overcome, and never take advantage of, the lack of necessary technological infrastructure, education and skills, as well as legal frameworks, particularly in LMICs, LDCs, LLDCs and SIDS, affecting communities.

Living in peaceful, just and interconnected societies

22. AI actors should play a participative and enabling role to ensure peaceful and just societies, which is based on an interconnected future for the benefit of all, consistent with human rights and fundamental freedoms. The value of living in peaceful and just societies points to the potential of AI systems to contribute throughout their life cycle to the interconnectedness of all living creatures with each other and with the natural environment.

23. The notion of humans being interconnected is based on the knowledge that every human belongs to a greater whole, which thrives when all its constituent parts are enabled to thrive. Living in peaceful, just and interconnected societies requires an organic, immediate, uncalculated bond of solidarity, characterized by a permanent search for

peaceful relations, tending towards care for others and the natural environment in the broadest sense of the term.

24. This value demands that peace, inclusiveness and justice, equity and interconnectedness should be promoted throughout the life cycle of AI systems, in so far as the processes of the life cycle of AI systems should not segregate, objectify or undermine freedom and autonomous decision-making as well as the safety of human beings and communities, divide and turn individuals and groups against each other, or threaten the coexistence between humans, other living beings and the natural environment.

15.4 Principles

(Chapter III.2) **Proportionality and do no harm**

25. It should be recognized that AI technologies do not necessarily, per se, ensure human and environmental and ecosystem flourishing. Furthermore, none of the processes related to the AI system life cycle shall exceed what is necessary to achieve legitimate aims or objectives and should be appropriate to the context. In the event of possible occurrence of any harm to human beings, human rights and fundamental freedoms, communities and society at large or the environment and ecosystems, the implementation of procedures for risk assessment and the adoption of measures in order to preclude the occurrence of such harm should be ensured.

26. The choice to use AI systems and which AI method to use should be justified in the following ways: (a) the AI method chosen should be appropriate and proportional to achieve a given legitimate aim; (b) the AI method chosen should not infringe upon the foundational values captured in this document, in particular, its use must not violate or abuse human rights; and (c) the AI method should be appropriate to the context and

should be based on rigorous scientific foundations. In scenarios where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life and death decisions, final human determination should apply. In particular, AI systems should not be used for social scoring or mass surveillance purposes.

Safety and security

27. Unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be avoided and should be addressed, prevented and eliminated throughout the life cycle of AI systems to ensure human, environmental and ecosystem safety and security. Safe and secure AI will be enabled by the development of sustainable, privacy-protective data access frameworks that foster better training and validation of AI models utilizing quality data.

Fairness and non-discrimination

28. AI actors should promote social justice and safeguard fairness and non-discrimination of any kind in compliance with international law. This implies an inclusive approach to ensuring that the benefits of AI technologies are available and accessible to all, taking into consideration the specific needs of different age groups, cultural systems, different language groups, persons with disabilities, girls and women, and disadvantaged, marginalized and vulnerable people or people in vulnerable situations. Member States should work to promote inclusive access for all, including local communities, to AI systems with locally relevant content and services, and with respect for multilingualism and cultural diversity. Member States should work to tackle digital divides and ensure inclusive access to and participation in the development of AI. At the national level, Member States should promote equity between rural and urban areas, and among all persons regardless of race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any

other grounds, in terms of access to and participation in the AI system life cycle. At the international level, the most technologically advanced countries have a responsibility of solidarity with the least advanced to ensure that the benefits of AI technologies are shared such that access to and participation in the AI system life cycle for the latter contributes to a fairer world order with regard to information, communication, culture, education, research and socio-economic and political stability.

29. AI actors should make all reasonable efforts to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system to ensure fairness of such systems. Effective remedy should be available against discrimination and biased algorithmic determination.

30. Furthermore, digital and knowledge divides within and between countries need to be addressed throughout an AI system life cycle, including in terms of access and quality of access to technology and data, in accordance with relevant national, regional and international legal frameworks, as well as in terms of connectivity, knowledge and skills and meaningful participation of the affected communities, such that every person is treated equitably.

Sustainability

31. The development of sustainable societies relies on the achievement of a complex set of objectives on a continuum of human, social, cultural, economic and environmental dimensions. The advent of AI technologies can either benefit sustainability objectives or hinder their realization, depending on how they are applied across countries with varying levels of development. The continuous assessment of the human, social, cultural, economic and environmental impact of AI technologies should therefore be carried out with full cognizance of the implications of AI technologies for sustainability as a set of constantly evolving goals

across a range of dimensions, such as currently identified in the Sustainable Development Goals (SDGs) of the United Nations.

Right to Privacy, and Data Protection

32. Privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of AI systems. It is important that data for AI systems be collected, used, shared, archived and deleted in ways that are consistent with international law and in line with the values and principles set forth in this Recommendation, while respecting relevant national, regional and international legal frameworks.

33. Adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach at the national or international level, protected by judicial systems, and ensured throughout the life cycle of AI systems. Data protection frameworks and any related mechanisms should take reference from international data protection principles and standards concerning the collection, use and disclosure of personal data and exercise of their rights by data subjects while ensuring a legitimate aim and a valid legal basis for the processing of personal data, including informed consent.

34. Algorithmic systems require adequate priacy impact assessments, which also include societal and ethical considerations of their use and an innovative use of the privacy by design approach. AI actors need to ensure that they are accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system.

Human oversight and determination

35. Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems, as well as in cases of remedy related to AI systems, to physical persons or to existing legal entities. Human oversight refers thus not only

to individual human oversight, but to inclusive public oversight, as appropriate.

36. It may be the case that sometimes humans would choose to rely on AI systems for reasons of efficacy, but the decision to cede control in limited contexts remains that of humans, as humans can resort to AI systems in decision-making and acting, but an AI system can never replace ultimate human responsibility and accountability. As a rule, life and death decisions should not be ceded to AI systems.

Transparency and explainability

37. The transparency and explainability of AI systems are often essential preconditions to ensure the respect, protection and promotion of human rights, fundamental freedoms and ethical principles. Transparency is necessary for relevant national and international liability regimes to work effectively. A lack of transparency could also undermine the possibility of effectively challenging decisions based on outcomes produced by AI systems and may thereby infringe the right to a fair trial and effective remedy, and limits the areas in which these systems can be legally used.

38. While efforts need to be made to increase transparency and explainability of AI systems, including those with extra-territorial impact, throughout their life cycle to support democratic governance, the level of transparency and explainability should always be appropriate to the context and impact, as there may be a need to balance between transparency and explainability and other principles such as privacy, safety and security. People should be fully informed when a decision is informed by or is made on the basis of AI algorithms, including when it affects their safety or human rights, and in those circumstances should have the opportunity to request explanatory information from the relevant AI actor or public sector institutions. In addition, individuals should be able to access the reasons for a decision affecting their rights and freedoms, and

have the option of making submissions to a designated staff member of the private sector company or public sector institution able to review and correct the decision. AI actors should inform users when a product or service is provided directly or with the assistance of AI systems in a proper and timely manner.

39. From a socio-technical lens, greater transparency contributes to more peaceful, just, democratic and inclusive societies. It allows for public scrutiny that can decrease corruption and discrimination, and can also help detect and prevent negative impacts on human rights. Transparency aims at providing appropriate information to the respective addressees to enable their understanding and foster trust. Specific to the AI system, transparency can enable people to understand how each stage of an AI system is put in place, appropriate to the context and sensitivity of the AI system. It may also include insight into factors that affect a specific prediction or decision, and whether or not appropriate assurances (such as safety or fairness measures) are in place. In cases of serious threats of adverse human rights impacts, transparency may also require the sharing of code or datasets.

40. Explainability refers to making intelligible and providing insight into the outcome of AI systems. The explainability of AI systems also refers to the understandability of the input, output and the functioning of each algorithmic building block and how it contributes to the outcome of the systems. Thus, explainability is closely related to transparency, as outcomes and sub-processes leading to outcomes should aim to be understandable and traceable, appropriate to the context. AI actors should commit to ensuring that the algorithms developed are explainable. In the case of AI applications that impact the end user in a way that is not temporary, easily reversible or otherwise low risk, it should be ensured that the meaningful explanation is provided with any decision that resulted in the action taken in order for the outcome to be considered transparent.

41. Transparency and explainability relate closely to adequate responsibility and accountability measures, as well as to the trustworthiness of AI systems.

Responsibility and accountability

42. AI actors and Member States should respect, protect and promote human rights and fundamental freedoms, and should also promote the protection of the environment and ecosystems, assuming their respective ethical and legal responsibility, in accordance with national and international law, in particular Member States' human rights obligations, and ethical guidance throughout the life cycle of AI systems, including with respect to AI actors within their effective territory and control. The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system.

43. Appropriate oversight, impact assessment, audit and due diligence mechanisms, including whistle-blowers' protection, should be developed to ensure accountability for AI systems and their impact throughout their life cycle. Both technical and institutional designs should ensure auditability and traceability of (the working of) AI systems in particular to address any conflicts with human rights norms and standards and threats to environmental and ecosystem well-being.

Awareness and literacy

44. Public awareness and understanding of AI technologies and the value of data should be promoted through open and accessible education, civic engagement, digital skills and AI ethics training, media and information literacy and training led jointly by governments, intergovernmental organizations, civil society, academia, the media, community leaders and the private sector, and considering the existing linguistic, social and cultural diversity, to ensure effective public participation so that all

members of society can take informed decisions about their use of AI systems and be protected from undue influence.

45. Learning about the impact of AI systems should include learning about, through and for human rights and fundamental freedoms, meaning that the approach and understanding of AI systems should be grounded by their impact on human rights and access to rights, as well as on the environment and ecosystems.

Multi-stakeholder and adaptive governance and collaboration

46. International law and national sovereignty must be respected in the use of data. That means that States, complying with international law, can regulate the data generated within or passing through their territories, and take measures towards effective regulation of data, including data protection, based on respect for the right to privacy in accordance with international law and other human rights norms and standards.

47. Participation of different stakeholders throughout the AI system life cycle is necessary for inclusive approaches to AI governance, enabling the benefits to be shared by all, and to contribute to sustainable development. Stakeholders include but are not limited to governments, intergovernmental organizations, the technical community, civil society, researchers and academia, media, education, policy-makers, private sector companies, human rights institutions and equality bodies, anti-discrimination monitoring bodies, and groups for youth and children. The adoption of open standards and interoperability to facilitate collaboration should be in place. Measures should be adopted to take into account shifts in technologies, the emergence of new groups of stakeholders, and to allow for meaningful participation by marginalized groups, communities and individuals and, where relevant, in the case of Indigenous Peoples, respect for the self-governance of their data.

16

UNESCO AI COMPETENCY FRAMEWORK FOR STUDENTS

16.1 Introduction

Unesco published its "AI Competency Framework for Students" in 2024[423]. Students and teachers are AI users, co-creators and present or future decision-makers, regulators and co-responsible for governance. Unesco positions its framework in the frame of the 2030 SDG agenda. Unesco as the UN agency for education has its priority on SDG goal 4 which aims "to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". The framework for students was developed along with the Unesco Competency Framework for Teachers.[424] The following text is an excerpt of the full document, covering the Executive Summary (p.1), Introduction (chapter 1, pp.12-13) and Key Principles (chapter 2, pp.14-17).

The Editors

---

[423] Unesco, *AI Competency Framework for Students,* Paris 2024, 80 pages, free download https://unesdoc.unesco.org/ark:/48223/pf0000391105.

[424] Unesco, *AI Competency Framework for Teachers,* Paris 2024, 52 pages, free download https://unesdoc.unesco.org/ark:/48223/pf0000391104.

16.2 Short summary

Preparing students to be responsible and creative citizens in the era of AI

Artificial intelligence (AI) is increasingly integral to our lives, necessitating proactive education systems to prepare students to be responsible users and co-creators of AI. Integrating AI learning objectives into official school curricula is crucial for students globally to engage safely and meaningfully with AI. The UNESCO AI competency framework for students aims to help educators in this integration, outlining 12 competencies across four dimensions: Human-centred mindset, Ethics of AI, AI techniques and applications, and AI system design. These competencies span three progression levels: Understand, Apply, and Create. The framework details curricular goals and domain-specific pedagogical methodologies. Grounded in a vision of students as AI co-creators and responsible citizens, the framework emphasizes critical judgement of AI solutions, awareness of citizenship responsibilities in the era of AI, foundational AI knowledge for lifelong learning, and inclusive, sustainable AI design.

16.3 Introduction: Why an AI framework?

1.1 Why an AI competency framework for students? The rapid iterations and proliferation of artificial intelligence (AI) across all aspects of life and all sectors are posing new challenges regarding the nature of machine intelligence, the collection and use of personal data, the role of humans and machines in decision-making, and the impact of AI on social and environmental sustainability. It is essential that education systems prepare students not only with the knowledge and skills to use AI, but also with insight into the potential impact of technology on societies and the environment at large. Given the transformative potential of AI for human societies, it is crucial to equip students with the values, knowledge

and skills needed for the effective use and active co-creation of AI. Education, as a public sector, cannot be reduced to a testing ground for the passive adoption of AI. The role of the education sector is not only to prepare students to adapt to a society that is increasingly being transformed by AI technologies; it also has a key role to play in empowering young people to help co-create sustainable futures by rebalancing our relationships, not only with others, but also with technology and the environment. By defining the core competencies that students are likely to require as we move deeper into the AI era, the ultimate aim of this AI competency framework for students (AI CFS) is to help shape responsible and creative citizens that can co-create these desirable futures. Governments acknowledged the urgent need to develop AI literacy and more advanced AI competencies as early as 2019, when they adopted the UNESCO Beijing Consensus on AI and Education. Indeed, the Beijing Consensus underlined the need to equip people with AI literacy across all layers of society. However, according to a recent survey conducted across 190 countries, only some 15 countries were found to be developing or implementing AI curricula in school education (UNESCO, 2022b). The survey also found that there was wide variation in how countries defined AI literacy, skills and competency. The results of the survey therefore underscored the urgency of developing a harmonized approach to integrating AI-related teaching and learning content in school curricula. Far too often, the definition of AI competencies for students is influenced by training designed and/or provided by private companies, which tends to focus on technical skills to operate profit-driven AI platforms. Such approaches seldom engage with the broader critical issues of the implications of AI for learning and citizenship, more broadly. There is currently a void in too many education systems when it comes to public-approved frameworks for introducing AI-related content and methods to educational curricula. One of the challenges that public education

systems are facing in filling this void is the lack of an international reference framework on AI competencies for students. Such an international reference framework can inform the design of national/local AI competency frameworks for students that

AI competency framework for students – Chapter 1: Introduction13promote a critical and ethical approach to AI tools, as well as develop the foundational knowledge required for their effective and meaningful use in education. The aim of this AI CFS is to fill this void. AI technology is a rapidly moving target. It is therefore critical to ensure that all students have a core set of knowledge, skills and values for interacting ethically and effectively with AI in the present. This foundation can enable students to utilize future iterations of AI technology in an appropriate and human-centred manner. The AI CFS supports educational authorities to respond to these needs by defining a core set of competencies for students that fall under four aspects: Human-centred mindset; Ethics of AI; AI techniques and applications; and AI system design. These four aspects are articulated at three levels of progression or mastery (understanding, application and creation), resulting in a total of twelve competency blocks. For each of these competency blocks, the AI CFS proposes detailed specifications on relevant pedagogical methodologies and strategies for the planning and provision of AI-related curricular content.

1.2 Purpose and target audience. The AI CFS aims to serve as a guide for public education systems to build the competencies required of all students and citizens for the effective implementation of national AI strategies and the building of inclusive, just and sustainable futures in this new technological era. More specifically, the AI CFS: (1) provides a global reference framework on the core set of AI competencies for students to inform the design of national or institutional AI competency frameworks; (2) specifies typical attitudinal and behavioural

performance relating to the key aspects of AI competencies at different levels of mastery to help design AI-related curricular content for school students; and (3) recommends an open-ended roadmap to help plan the learning sequence of AI curricula across grade levels. As a global reference framework, the AI CFS is to be tailored to the diverse readiness levels of local education systems in terms of curricula, the enabling learning environment for teaching AI, preparedness of teachers, and the prior knowledge and capacities of specific groups of students. The AI CFS is aimed principally at policy-makers, curriculum developers, providers of education programmes on AI for students, school leaders, teachers and educational experts.

16.4 Key principles

2.1 Fostering a critical approach to AI Critical thinking is a fundamental skill that students need to meaningfully engage with AI as learners, users and creators. Students also have the responsibility to determine what types of AI should be developed and how they should be used to drive human societies towards inclusive, environmentally sound, shared futures. School students need to be supported to become active co-creators of AI, as well as potential leaders who will define further iterations of AI and its interactions with human society for present and future generations. To support this vision, the AI CFS is designed to foster a critical approach to AI by engaging students with fundamental questions, such as: is AI poised to help solve real-world challenges faced by humans, or does it pose insurmountable threats to humans? Are adverse impacts on climate of training and using AI disproportionate to its anticipated benefits? What social, economic, political and demographic impacts of the use of AI should be carefully reviewed? The AI-driven transformation across development sectors has profound implications for human agency,

human interactions, social equity, economic inclusiveness, and environmental sustainability. Thus, in the first place, school students are expected to be conscious and knowledgeable of the advantages and limitations of existing affordances of AI. The pre-condition for responsible use consists in students' abilities to detect the trustworthiness and proportionality of AI tools. The AI CFS aims to prepare students with the values, knowledge and skills necessary to critically examine the proportionality of AI from an ethical perspective. This includes examining and understanding its impact on human agency, social inclusion and equity, institutional and individual security, cultural and linguistic diversity, the construction and expression of plural opinions, as well as on the environment and on ecosystems. Students are expected to move beyond the misconception that AI is a solution to everything. Rather, they are to become conscious decision-makers on when AI systems and applications should, or should not, be used; what problems they may or may not solve; and when and how AI should be designed and used as one part of a wider solution. The AI CFS aims to nurture students' aspirations to apply and design AI tools to serve meaningful specific purposes or to address real-world challenges and promote sustainable development. Societies are moving into the era of AI at different paces, but students everywhere are, or will be, citizens in contexts characterized by widespread AI integration. They will not only have to comply with legal regulations and ethical principles, but, as citizens, they will also have to contribute to the adaptation of AI standards and regulations. The framework therefore highlights the importance of supporting students to become responsible and ethical users of, as well as contributors to, AI. It engages students to reflect on key controversies surrounding AI, internalize ethical principles, and become familiar with related regulations.

AI competency framework for students – Chapter 2: Key principles. The AI CFS sets out a forward-looking vision of the type of citizenship

required by societies increasingly shaped by AI. It proposes that students be challenged and enabled to make meaningful use of AI for self-actualization; to evaluate its social, economic and environmental impacts; and to contribute, at a level appropriate for their age or grade, to the development of AI regulations, helping to shape our relationship with technology in society at large.

2.2 Prioritizing human-centred interaction with AI. In the era of AI, interaction between humans and AI systems and applications will become an essential constituent element of public service, production and commerce, social practice, learning, and daily life. Establishing the competencies needed to understand and ensure human-centred interaction with AI in these domains is a priority for the AI CFS. UNESCO's human-centric approach advocates that the design and use of AI should serve the development of human capabilities, protect human dignity and agency, and promote justice and sustainability throughout the entire AI life cycle and all possible human–AI interaction loops. Such an approach must be guided by human rights principles and respect for the linguistic and cultural diversity that defines the knowledge commons. A human-centred approach also requires that AI be used in ways that ensure transparency and explainability, as well as human control and accountability. As AI becomes increasingly sophisticated and more widely used, a key danger is its potential to undermine human agency and compromise the development of human intellectual skills. While AI can be used to challenge and extend human thought, it should not be allowed to usurp or replace critical thinking. The protection and enhancement of human agency should, therefore, always be a core principle in the design of AI curricula and education programmes. The AI CFS aims to support students to understand the types of data that AI may collect from them, the methods with which the data may be used to train AI models, and the impact that the data cycle may have on their privacy and wider lives. It seeks to stimulate

students' intrinsic motivation to grow and learn as individuals and to reinforce their autonomy in contexts in which sophisticated AI systems are increasingly being integrated. Critical AI competencies, as proposed in this framework, can also guide students to understand the unique value of social interaction and of the creative works produced by humans that should not be replaced by AI outputs. By developing competencies for human-centred engagement with AI, the framework aims to prevent students from becoming addicted to or dependent on AI, and to foster behaviours that maintain human accountability for high-stakes decisions.

2.3 Encouraging environmentally sustainable AIAs co-creators and potential leaders of the next generations of AI technology, students need to have a critical understanding of the adverse environmental impact of profit-driven approaches to the design, training and deployment of AI models. Education systems bear the responsibility of ensuring that students understand carbon emissions,

AI competency framework for students –Chapter 2: Key principles analyse the root causes of climate change, and act judiciously to protect the climate and the environment. In the race to produce increasingly powerful AI models, environmental sustainability is often considered to be of secondary importance. In some instances, it has even been intentionally obscured by claims that AI holds the promise of solving climate change. As global leaders and policy-makers work to consider regulations around the consumption of energy and the protection of the environment, it is imperative that students understand how the training of AI models is contributing to the destruction of the natural environment. Learning about AI should empower them to urgently explore more climate-friendly approaches to the design, training and use of AI models. The AI CFS attends to this by guiding students to design and implement project-based learning activities on the environmental impacts of AI use and training,

prompting students to investigate potential solutions to mitigate these impacts.

2.4 Promoting inclusivity in AI competency development. Access to AI and AI competencies represent the two sides of citizens' basic rights in today's world. All students should have inclusive access to the environments required for learning about AI at the basic level, and they should be supported to learn how to embed the principle of inclusivity into the design of AI and be prepared to contribute to an inclusive AI society. When defining AI competencies, school students should be provided with opportunities to understand and apply the principle of inclusivity across the AI life cycle. This covers the selection of representative data, the choice of bias-agnostic algorithms and anti-discrimination training methods, the design of accessible functionalities, testing for the inclusiveness of AI outputs, and impact assessment of the use of AI on social inclusion. With regard to AI system design, students can deepen their understanding and application skills to assess the needs of users with different abilities as well as those from diverse linguistic and cultural backgrounds. In selecting the models and categories of technologies as vectors of AI-related teaching and learning, care is needed to avoid favouring certain demographics over others.

When recommending specific AI tools for educational purposes, rigorous public validation mechanisms must be applied to avoid algorithms with bias(es) related to gender, ability, socio-economic status, language, ethnicity and/or culture. AI tools that are designed to support individuals with disabilities and promote linguistic and cultural diversity should be given priority. Where such validation mechanisms are unavailable, the recommendation of specific AI tools for use at scale should be avoided. Turning to delivery of the curriculum, specific measures can be outlined to provide basic enabling conditions for the implementation of the AI CFS-based curriculum.

While AI frameworks or educational programmes should be designed to be applicable to all students, including those who live in low-tech settings, engagement with AI without access to the internet and AI tools will limit the scope and mastery level of AI competencies. Governments should commit to promoting inclusive access to basic internet connectivity, updated digital devices, open-source or affordable AI.

17

OECD RECOMMENDATION ON ARTIFICIAL INTELLIGENCE

17.1 Introduction

The OECD Council adopted the "Recommendation on Artificial Intelligence"[425] in 2024 as guidelines for all OECD member countries. Section 1 formulates "principles for responsible stewardship of trustworthy AI" on transparency, fairness, accountability, privacy and human rights. In 2019, OECD already published similar "Principles on AI", adopted by member countries.

Section 2 deals with "National policies and international co-operation for trustworthy AI". The following text is the full document of the Recommendation, but without the introductory background information.

The Editors

[425] OECD, *Recommendation of the Council on Artificial Intelligence,* OECD/Legal/0449, OECD 2024. Free download http://legalinstruments.oecd.org.

17.2 Recommendation

THE COUNCIL,

HAVING REGARD to Article 5 b) of the Convention on the Organisation for Economic Co-operation and Development of 14 December 1960;

HAVING REGARD to the OECD Guidelines for Multinational Enterprises [OECD/LEGAL/0144];

Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [OECD/LEGAL/0188]; Recommendation of the Council concerning Guidelines for Cryptography Policy [OECD/LEGAL/0289]; Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information [OECD/LEGAL/0362]; Recommendation of the Council on Digital Security Risk Management for Economic and Social Prosperity [OECD/LEGAL/0415]; Recommendation of the Council on Consumer Protection in E-commerce [OECD/LEGAL/0422]; Declaration on the Digital Economy: Innovation, Growth and Social Prosperity (Cancún Declaration) [OECD/LEGAL/0426]; Declaration on Strengthening SMEs and Entrepreneurship for Productivity and Inclusive Growth [OECD/LEGAL/0439]; as well as the 2016 Ministerial Statement on Building more Resilient and Inclusive Labour Markets, adopted at the OECD Labour and Employment Ministerial Meeting;

HAVING REGARD to the Sustainable Development Goals set out in the 2030 Agenda for Sustainable Development adopted by the United Nations General Assembly (A/RES/70/1) as well as the 1948 Universal Declaration of Human Rights;

HAVING REGARD to the important work being carried out on artificial intelligence (hereafter, "AI") in other international governmental and non-governmental fora;

RECOGNISING that AI has pervasive, far-reaching and global implications that are transforming societies, economic sectors and the world of work, and are likely to increasingly do so in the future;

RECOGNISING that AI has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges;

RECOGNISING that, at the same time, these transformations may have disparate effects within, and between societies and economies, notably regarding economic shifts, competition, transitions in the labour market, inequalities, and implications for democracy and human rights, privacy and data protection, and digital security;

RECOGNISING that trust is a key enabler of digital transformation; that, although the nature of future AI applications and their implications may be hard to foresee, the trustworthiness of AI systems is a key factor for the diffusion and adoption of AI; and that a well-informed whole-of-society public debate is necessary for capturing the beneficial potential of the technology, while limiting the risks associated with it;

UNDERLINING that certain existing national and international legal, regulatory and policy frameworks already have relevance to AI, including those related to human rights, consumer and personal data protection, intellectual property rights, responsible business conduct, and competition, while noting that the appropriateness of some frameworks may need to be assessed and new approaches developed;

RECOGNISING that given the rapid development and implementation of AI, there is a need for a stable policy environment that promotes a human-centric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and that applies to all stakeholders according to their role and the context;

CONSIDERING that embracing the opportunities offered, and addressing the challenges raised, by AI applications, and empowering stakeholders to engage is essential to fostering adoption of trustworthy AI in society, and to turning AI trustworthiness into a competitive parameter in the global marketplace;

On the proposal of the Committee on Digital Economy Policy:

I. AGREES that for the purpose of this Recommendation the following terms should be understood as follows:

– *AI system*: 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.

– *AI system lifecycle*: AI system lifecycle phases involve: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.

– *AI knowledge*: AI knowledge refers to the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle.

– *AI actors*: AI actors are those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI.

– *Stakeholders*: Stakeholders encompass all organisations and individuals involved in, or affected by, AI systems, directly or indirectly. AI actors are a subset of stakeholders.

17.2.1 Section 1: Principles for responsible stewardship of trustworthy AI

II. RECOMMENDS that Members and non-Members adhering to this Recommendation (hereafter the "Adherents") promote and implement the following principles for responsible stewardship of trustworthy AI, which are relevant to all stakeholders.

III. CALLS ON all AI actors to promote and implement, according to their respective roles, the following Principles for responsible stewardship of trustworthy AI.

IV. UNDERLINES that the following principles are complementary and should be considered as a whole.

1.1. Inclusive growth, sustainable development and well-being

Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

1.2. Human-centred values and fairness

a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.

b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

1.3. Transparency and explainability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

i. to foster a general understanding of AI systems,

ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,

iii. to enable those affected by an AI system to understand the outcome, and,

iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

1.4. Robustness, security and safety

a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.

b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

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

1.5. Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

17.2.2 Section 2: National policies and international co-operation for trustworthy AI

V. RECOMMENDS that Adherents implement the following recommendations, consistent with the principles in section 1, in their national policies and international co-operation, with special attention to small and medium-sized enterprises (SMEs).

2.1. Investing in AI research and development

a) Governments should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues.

b) Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.

2.2. Fostering a digital ecosystem for AI

Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.

2.3. Shaping an enabling policy environment for AI

a) Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate.

b) Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.

2.4. Building human capacity and preparing for labour market transformation

a) Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills.

b) Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programmes along the working life, support for those affected by displacement, and access to new opportunities in the labour market.

c) Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits from AI are broadly and fairly shared.

2.5. International co-operation for trustworthy AI

a) Governments, including developing countries and with stakeholders, should actively co-operate to advance these principles and to progress on responsible stewardship of trustworthy AI.

b) Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate.

They should encourage international, cross-sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI.

c) Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI.

d) Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.

VI. INVITES the Secretary-General and Adherents to disseminate this Recommendation.

VII. INVITES non-Adherents to take due account of, and adhere to, this Recommendation.

VIII. INSTRUCTS the Committee on Digital Economy Policy:

a) to continue its important work on artificial intelligence building on this Recommendation and taking into account work in other international fora, and to further develop the measurement framework for evidence-based AI policies;

b) to develop and iterate further practical guidance on the implementation of this Recommendation, and to report to the Council on progress made no later than end December 2019;

c) to provide a forum for exchanging information on AI policy and activities including experience with the implementation of this Recommendation, and to foster multi-stakeholder and interdisciplinary dialogue to promote trust in and adoption of AI; and

d) to monitor, in consultation with other relevant Committees, the implementation of this Recommendation and report thereon to the Council no later than five years following its adoption and regularly thereafter.

18

EUROPEAN UNION: AI ACT OVERVIEW

18.1 Introduction

The European Union AI Act (in short EU AI ACT) was adopted in March 2024 by the EU Parliament and in May 2024 by the EU Council. It will be fully applicable 24 months after entry into force. It is called the world's first comprehensive AI law (even though other conventions and rules, listed in this book, are also binding). The following text is an excerpt of the overview, which is available online[426] The Act is a key, new policy document of the European countries. It introduces a risk-based approach, categorizing AI systems into four levels of risk: unacceptable, high, limited, and minimal. The Act focuses on ensuring transparency, accountability, and safety, with stringent requirements for high-risk AI applications, such as those in critical infrastructure or employment. As explained by the Council of Europe, banned or prohibited AI systems include, for example, cognitive behavioural manipulation, and social scoring and the use of AI for predictive policing based on profiling and systems that use biometric data to categorise people according to specific

[426] https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

categories such as race, religion, or sexual orientation. Its goal is to foster trust in AI while encouraging innovation and protecting fundamental rights.

The Editors

18.2 The AI Act classifies AI according to its risk:

- Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI).
- Most of the text addresses high-risk AI systems, which are regulated.
- A smaller section handles limited risk AI systems, subject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).
- Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filters – at least in 2021; this is changing with generative AI).

18.3 The majority of obligations fall on providers (developers) of high-risk AI systems

- Those that intend to place on the market or put into service high-risk AI systems in the EU, regardless of whether they are based in the EU or a third country.
- And also third country providers where the high risk AI system's output is used in the EU.

Deployers are natural or legal persons that deploy an AI system in a professional capacity, not affected end-users.

- Deployers of high-risk AI systems have some obligations, though less than providers (developers).
- This applies to deployers located in the EU, and third country users where the AI system's output is used in the EU.

General purpose AI (GPAI):

- All GPAI model providers must provide technical documentation, instructions for use, comply with the Copyright Directive, and publish a summary about the content used for training.
- Free and open licence GPAI model providers only need to comply with copyright and publish the training data summary, unless they present a systemic risk.
- All providers of GPAI models that present a systemic risk – open or closed – must also conduct model evaluations, adversarial testing, track and report serious incidents and ensure cybersecurity protections.

18.4 Prohibited AI systems (Chapter II, Art. 5)

AI systems:

- deploying **subliminal, manipulative, or deceptive techniques** to distort behaviour and impair informed decision-making, causing significant harm.
- **exploiting vulnerabilities** related to age, disability, or socio-economic circumstances to distort behaviour, causing significant harm.
- **social scoring**, i.e., evaluating or classifying individuals or groups based on social behaviour or personal traits, causing detrimental or unfavourable treatment of those people.
- **assessing the risk of an individual committing criminal offenses** solely based on profiling or personality traits, except

when used to augment human assessments based on objective, verifiable facts directly linked to criminal activity.

- **compiling facial recognition databases** by untargeted scraping of facial images from the internet or CCTV footage.
- **inferring emotions in workplaces or educational institutions**, except for medical or safety reasons.
- **biometric categorisation systems** inferring sensitive attributes (race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation), except labelling or filtering of lawfully acquired biometric datasets or when law enforcement categorises biometric data.
- **'real-time' remote biometric identification (RBI) in publicly accessible spaces for law enforcement**, except when:
    - targeted searching for missing persons, abduction victims, and people who have been human trafficked or sexually exploited;
    - preventing specific, substantial and imminent threat to life or physical safety, or foreseeable terrorist attack; or
    - identifying suspects in serious crimes (e.g., murder, rape, armed robbery, narcotic and illegal weapons trafficking, organised crime, and environmental crime, etc.).
        - Using AI-enabled real-time RBI is only allowed *when not using the tool would cause harm,* particularly regarding the seriousness, probability and scale of such harm, and must account for affected persons' rights and freedoms.
        - Before deployment, police must complete a *fundamental rights impact assessment and register the system in the EU database,* though, in duly justified cases of urgency, deployment can commence without

registration, provided that it is registered later without undue delay.

- Before deployment, they also must obtain *authorisation from a judicial authority or independent administrative authority*[427], though, in duly justified cases of urgency, deployment can commence without authorisation, provided that authorisation is requested within 24 hours. If authorisation is rejected, deployment must cease immediately, deleting all data, results, and outputs.

18.5 High risk AI systems (Chapter III)

18.5.1 Classification rules for high-risk AI systems (Art. 6)

High risk AI systems are those:

- used as a safety component or a product covered by EU laws in Annex I *AND* required to undergo a third-party conformity assessment under those Annex I laws; *OR*
- those under Annex III use cases (below), except if:
  - the AI system performs a narrow procedural task;
  - improves the result of a previously completed human activity;
  - detects decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment without proper human review; or
  - performs a preparatory task to an assessment relevant for the purpose of the use cases listed in Annex III.
- The Commission can add or modify the above conditions through delegated acts if there is concrete evidence that an AI system

[427] Independent administrative authorities may be subject to greater political influence than judicial authorities (Hacker, 2024).

falling under Annex III does not pose a significant risk to health, safety and fundamental rights. They can also delete any of the conditions if there is concrete evidence that this is needed to protect people.

- AI systems are always considered high-risk if it profiles individuals, i.e. automated processing of personal data to assess various aspects of a person's life, such as work performance, economic situation, health, preferences, interests, reliability, behaviour, location or movement.
- Providers that believe their AI system, which fails under Annex III, is not high-risk, must document such an assessment before placing it on the market or putting it into service.
- 18 months after entry into force, the Commission will provide guidance on determining if an AI system is high risk, with list of practical examples of high-risk and non-high risk use cases.

18.5.2 Requirements for providers of high-risk AI systems (Art. 8-17)

High risk AI providers must:

- Establish a *risk management system* throughout the high risk AI system's lifecycle;
- Conduct *data governance*, ensuring that training, validation and testing datasets are relevant, sufficiently representative and, to the best extent possible, free of errors and complete according to the intended purpose.
- Draw up *technical documentation* to demonstrate compliance and provide authorities with the information to assess that compliance.
- Design their high risk AI system for *record-keeping* to enable it to automatically record events relevant for identifying national level risks and substantial modifications throughout the system's lifecycle.

- Provide *instructions for use* to downstream deployers to enable the latter's compliance.
- Design their high risk AI system to allow deployers to implement *human oversight.*
- Design their high risk AI system to achieve appropriate levels of **ac***curacy, robustness, and cybersecurity.*
- Establish a *quality management system* to ensure compliance.

18.6 General purpose AI (GPAI) (Chapter V)

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. This does not cover AI models that are used before release on the market for research, development and prototyping activities.

GPAI system means an AI system which is based on a general purpose AI model, that has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems. GPAI systems may be used as high risk AI systems or integrated into them. GPAI system providers should cooperate with such high risk AI system providers to enable the latter's compliance.

18.6.1 All providers of GPAI models must (Art. 53):

- Draw up **technical documentation**, including training and testing process and evaluation results.
- Draw up **information and documentation to supply to downstream providers** that intend to integrate the GPAI model into their own AI system in order that the latter understands capabilities and limitations and is enabled to comply.

- Establish a policy to **respect the Copyright Directive.**
- Publish a **sufficiently detailed summary about the content used for training** the GPAI model.

Free and open licence GPAI models – whose parameters, including weights, model architecture and model usage are publicly available, allowing for access, usage, modification and distribution of the model – only have to comply with the latter two obligations above, unless the free and open licence GPAI model is systemic.

GPAI models are considered systemic when the cumulative amount of compute used for its training is greater than 10^25 floating point operations per second (FLOPS) (Art. 51). Providers must notify the Commission if their model meets this criterion within 2 weeks (Art. 52). The provider may present arguments that, despite meeting the criteria, their model does not present systemic risks. The Commission may decide on its own, or via a qualified alert from the scientific panel of independent experts, that a model has high impact capabilities, rendering it systemic.

In addition to the four obligations above, providers of GPAI models with systemic risk must also (Art. 55):

- Perform *model evaluations*, including conducting and documenting *adversarial testing* to identify and mitigate systemic risk.
- *Assess and mitigate possible systemic risks*, including their sources.
- *Track, document and report serious incidents* and possible corrective measures to the AI Office and relevant national competent authorities without undue delay.
- Ensure an adequate level of *cybersecurity protection.*

All GPAI model providers may demonstrate compliance with their obligations if they voluntarily adhere to codes of practice until European harmonised standards are published, compliance with which will lead to a presumption of conformity (Art. 56). Providers that don't adhere to

codes of practice must demonstrate *alternative adequate means of compliance* for Commission approval.

18.6.2 Codes of practice (Art. 56)

- Will account for international approaches.
- Will cover but not necessarily limited to the above obligations, particularly the relevant information to include in technical documentation for authorities and downstream providers, identification of the type and nature of systemic risks and their sources, and the modalities of risk management accounting for specific challenges in addressing risks due to the way they may emerge and materialise throughout the value chain.
- AI Office may invite GPAI model providers, relevant national competent authorities to participate in drawing up the codes, while civil society, industry, academia, downstream providers and independent experts may support the process.

18.7 Governance (Chapter VI)

- The AI Office will be established, sitting within the Commission, to monitor the effective implementation and compliance of GPAI model providers (Art. 64).
- Downstream providers can lodge a complaint regarding the upstream providers infringement to the AI Office (Art. 89).
- The AI Office may conduct evaluations of the GPAI model to (Art. 92):
    - assess compliance where the information gathered under its powers to request information is insufficient.

- Investigate systemic risks, particularly following a qualified report from the scientific panel of independent experts (Art. 90).

18.8 Timelines

- After entry into force, the AI Act will apply by the following deadlines:
  - 6 months for prohibited AI systems.
  - 12 months for GPAI.
  - 24 months for high risk AI systems under Annex III.
  - 36 months for high risk AI systems under Annex I.
- Codes of practice must be ready 9 months after entry into force.

19

COUNCIL OF EUROPE: FRAMEWORK CONVENTION ON ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS, DEMOCRACY AND THE RULE OF LAW

19.1 Introduction

The "Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law" was adopted by the member states of the Council of Europe on 5 September 2024.[428] Its specificity is that it shows "that human rights, democracy and the rule of law are inherently interwoven" and need to be seen together for governing and regulating AI. The convention deals specifically with manifold risks of AI and their possible remedies. Risks

The Editors

[428] Council of Europe: Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Vilnius 5 Sept 2024, Treaty Series - No. 225. Free download: https://rm.coe.int/1680afae3c

19.2 Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law

19.2.0 Preamble

The member States of the Council of Europe and the other signatories hereto,

Considering that the aim of the Council of Europe is to achieve greater unity between its members, based in particular on the respect for human rights, democracy and the rule of law;

Recognising the value of fostering co-operation between the Parties to this Convention and of extending such co-operation to other States that share the same values;

Conscious of the accelerating developments in science and technology and the profound changes brought about through activities within the lifecycle of artificial intelligence systems, which have the potential to promote human prosperity as well as individual and societal well-being, sustainable development, gender equality and the empowerment of all women and girls, as well as other important goals and interests, by enhancing progress and innovation;

Recognising that activities within the lifecycle of artificial intelligence systems may offer unprecedented opportunities to protect and promote human rights, democracy and the rule of law;

Concerned that certain activities within the lifecycle of artificial intelligence systems may undermine human dignity and individual autonomy, human rights, democracy and the rule of law;

Concerned about the risks of discrimination in digital contexts, particularly those involving artificial intelligence systems, and their potential effect of creating or aggravating inequalities, including those experienced by women and individuals in vulnerable situations, regarding the

enjoyment of their human rights and their full, equal and effective participation in economic, social, cultural and political affairs;

Concerned by the misuse of artificial intelligence systems and opposing the use of such systems for repressive purposes in violation of international human rights law, including through arbitrary or unlawful surveillance and censorship practices that erode privacy and individual autonomy;

Conscious of the fact that human rights, democracy and the rule of law are inherently interwoven;

Convinced of the need to establish, as a matter of priority, a globally applicable legal framework setting out common general principles and rules governing the activities within the lifecycle of artificial intelligence systems that effectively preserves shared values and harnesses the benefits of artificial intelligence for the promotion of these values in a manner conducive to responsible innovation;

Recognising the need to promote digital literacy, knowledge about, and trust in the design, development, use and decommissioning of artificial intelligence systems;

Recognising the framework character of this Convention, which may be supplemented by further instruments to address specific issues relating to the activities within the lifecycle of artificial intelligence systems;

Underlining that this Convention is intended to address specific challenges which arise throughout the lifecycle of artificial intelligence systems and encourage the consideration of the wider risks and impacts related to these technologies including, but not limited to, human health and the environment, and socio-economic aspects, such as employment and labour;

Noting relevant efforts to advance international understanding and co-operation on artificial intelligence by other international and supranational organisations and fora;

Mindful of applicable international human rights instruments, such as the 1948 Universal Declaration of Human Rights, the 1950 Convention for the Protection of Human Rights and Fundamental Freedoms (ETS No. 5), the 1966 International Covenant on Civil and Political Rights, the 1966 International Covenant on Economic, Social and Cultural Rights, the 1961 European Social Charter (ETS No. 35), as well as their respective protocols, and the 1996 European Social Charter (Revised) (ETS No. 163);

Mindful also of the 1989 United Nations Convention on the Rights of the Child and the 2006 United Nations Convention on the Rights of Persons with Disabilities;

Mindful also of the privacy rights of individuals and the protection of personal data, as applicable and conferred, for example, by the 1981 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108) and its protocols;

Affirming the commitment of the Parties to protecting human rights, democracy and the rule of law, and fostering trustworthiness of artificial intelligence systems through this Convention,

Have agreed as follows:

19.2.1 Chapter I – General provisions

Article 1 – Object and purpose

1. The provisions of this Convention aim to ensure that activities within the lifecycle of artificial intelligence systems are fully consistent with human rights, democracy and the rule of law.

2. Each Party shall adopt or maintain appropriate legislative, administrative or other measures to give effect to the provisions set out in this Convention. These measures shall be graduated and differentiated as may be necessary in view of the severity and probability of the occurrence of adverse impacts on human rights, democracy and the rule of law throughout the lifecycle of artificial intelligence systems. This may include

specific or horizontal measures that apply irrespective of the type of technology used.

3. In order to ensure effective implementation of its provisions by the Parties, this Convention establishes a follow-up mechanism and provides for international co-operation.

Article 2 – Definition of artificial intelligence systems

For the purposes of this Convention, "artificial intelligence system" means 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.

Article 3 – Scope

1. The scope of this Convention covers the activities within the lifecycle of artificial intelligence systems that have the potential to interfere with human rights, democracy and the rule of law as follows:

a. Each Party shall apply this Convention to the activities within the lifecycle of artificial intelligence systems undertaken by public authorities, or private actors acting on their behalf.

b. Each Party shall address risks and impacts arising from activities within the lifecycle of artificial intelligence systems by private actors to the extent not covered in subparagraph a in a manner conforming with the object and purpose of this Convention.

Each Party shall specify in a declaration submitted to the Secretary General of the Council of Europe at the time of signature, or when depositing its instrument of ratification, acceptance, approval or accession, how it intends to implement this obligation, either by applying the principles and obligations set forth in Chapters II to VI of this Convention to activities of private actors or by taking other appropriate measures to

fulfil the obligation set out in this subparagraph. Parties may, at any time and in the same manner, amend their declarations.

When implementing the obligation under this subparagraph, a Party may not derogate from or limit the application of its international obligations undertaken to protect human rights, democracy and the rule of law.

2. A Party shall not be required to apply this Convention to activities within the lifecycle of artificial intelligence systems related to the protection of its national security interests, with the understanding that such activities are conducted in a manner consistent with applicable international law, including international human rights law obligations, and with respect for its democratic institutions and processes.

3. Without prejudice to Article 13 and Article 25, paragraph 2, this Convention shall not apply to research and development activities regarding artificial intelligence systems not yet made available for use, unless testing or similar activities are undertaken in such a way that they have the potential to interfere with human rights, democracy and the rule of law.

4. Matters relating to national defence do not fall within the scope of this Convention.

19.2.2 Chapter II – General obligations

Article 4 – Protection of human rights

Each Party shall adopt or maintain measures to ensure that the activities within the lifecycle of artificial intelligence systems are consistent with obligations to protect human rights, as enshrined in applicable international law and in its domestic law.

Article 5 – Integrity of democratic processes and respect for the rule of law

1. Each Party shall adopt or maintain measures that seek to ensure that artificial intelligence systems are not used to undermine the integrity, independence and effectiveness of democratic institutions and processes,

including the principle of the separation of powers, respect for judicial independence and access to justice.

2. Each Party shall adopt or maintain measures that seek to protect its democratic processes in the context of activities within the lifecycle of artificial intelligence systems, including individuals' fair access to and participation in public debate, as well as their ability to freely form opinions.

19.2.3 Chapter III – Principles related to activities within the lifecycle of artificial intelligence systems

Article 6 – General approach

This chapter sets forth general common principles that each Party shall implement in regard to artificial intelligence systems in a manner appropriate to its domestic legal system and the other obligations of this Convention.

Article 7 – Human dignity and individual autonomy

Each Party shall adopt or maintain measures to respect human dignity and individual autonomy in relation to activities within the lifecycle of artificial intelligence systems.

Article 8 – Transparency and oversight

Each Party shall adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place in respect of activities within the lifecycle of artificial intelligence systems, including with regard to the identification of content generated by artificial intelligence systems.

Article 9 – Accountability and responsibility

Each Party shall adopt or maintain measures to ensure accountability and responsibility for adverse impacts on human rights, democracy and the rule of law resulting from activities within the lifecycle of artificial intelligence systems.

Article 10 – Equality and non-discrimination

1. Each Party shall adopt or maintain measures with a view to ensuring that activities within the lifecycle of artificial intelligence systems respect equality, including gender equality, and the prohibition of discrimination, as provided under applicable international and domestic law.

2. Each Party undertakes to adopt or maintain measures aimed at overcoming inequalities to achieve fair, just and equitable outcomes, in line with its applicable domestic and international human rights obligations, in relation to activities within the lifecycle of artificial intelligence systems.

Article 11 – Privacy and personal data protection

Each Party shall adopt or maintain measures to ensure that, with regard to activities within the lifecycle of artificial intelligence systems:

a. privacy rights of individuals and their personal data are protected, including through applicable domestic and international laws, standards and frameworks; and

b. effective guarantees and safeguards have been put in place for individuals, in accordance with applicable domestic and international legal obligations.

Article 12 – Reliability

Each Party shall take, as appropriate, measures to promote the reliability of artificial intelligence systems and trust in their outputs, which could include requirements related to adequate quality and security throughout the lifecycle of artificial intelligence systems.

Article 13 – Safe innovation

With a view to fostering innovation while avoiding adverse impacts on human rights, democracy and the rule of law, each Party is called upon to enable, as appropriate, the establishment of controlled environments for developing, experimenting and testing artificial intelligence systems under the supervision of its competent authorities.

19.2.4 Chapter IV – Remedies

Article 14 – Remedies

1. Each Party shall, to the extent remedies are required by its international obligations and consistent with its domestic legal system, adopt or maintain measures to ensure the availability of accessible and effective remedies for violations of human rights resulting from the activities within the lifecycle of artificial intelligence systems.

2. With the aim of supporting paragraph 1 above, each Party shall adopt or maintain measures including:

a. measures to ensure that relevant information regarding artificial intelligence systems which have the potential to significantly affect human rights and their relevant usage is documented, provided to bodies authorised to access that information and, where appropriate and applicable, made available or communicated to affected persons;

b. measures to ensure that the information referred to in subparagraph a is sufficient for the affected persons to contest the decision(s) made or substantially informed by the use of the system, and, where relevant and appropriate, the use of the system itself; and

c. an effective possibility for persons concerned to lodge a complaint to competent authorities.

Article 15 – Procedural safeguards

1. Each Party shall ensure that, where an artificial intelligence system significantly impacts upon the enjoyment of human rights, effective procedural guarantees, safeguards and rights, in accordance with the applicable international and domestic law, are available to persons affected thereby.

2. Each Party shall seek to ensure that, as appropriate for the context, persons interacting with artificial intelligence systems are notified that they are interacting with such systems rather than with a human.

19.2.5 Chapter V – Assessment and mitigation of risks and adverse impacts

Article 16 – Risk and impact management framework

1. Each Party shall, taking into account the principles set forth in Chapter III, adopt or maintain measures for the identification, assessment, prevention and mitigation of risks posed by artificial intelligence systems by considering actual and potential impacts to human rights, democracy and the rule of law.

2. Such measures shall be graduated and differentiated, as appropriate, and:

a. take due account of the context and intended use of artificial intelligence systems, in particular as concerns risks to human rights, democracy, and the rule of law;

b. take due account of the severity and probability of potential impacts;

c. consider, where appropriate, the perspectives of relevant stakeholders, in particular persons whose rights may be impacted;

d. apply iteratively throughout the activities within the lifecycle of the artificial intelligence system;

e. include monitoring for risks and adverse impacts to human rights, democracy, and the rule of law;

f. include documentation of risks, actual and potential impacts, and the risk management approach; and

g. require, where appropriate, testing of artificial intelligence systems before making them available for first use and when they are significantly modified.

3. Each Party shall adopt or maintain measures that seek to ensure that adverse impacts of artificial intelligence systems to human rights, democracy, and the rule of law are adequately addressed. Such adverse impacts and measures to address them should be documented and inform the relevant risk management measures described in paragraph 2.

4. Each Party shall assess the need for a moratorium or ban or other appropriate measures in respect of certain uses of artificial intelligence systems where it considers such uses incompatible with the respect for human rights, the functioning of democracy or the rule of law.

19.2.6 Chapter VI – Implementation of the Convention

Article 17 – Non-discrimination

The implementation of the provisions of this Convention by the Parties shall be secured without discrimination on any ground, in accordance with their international human rights obligations.

Article 18 – Rights of persons with disabilities and of children

Each Party shall, in accordance with its domestic law and applicable international obligations, take due account of any specific needs and vulnerabilities in relation to respect for the rights of persons with disabilities and of children.

Article 19 – Public consultation

Each Party shall seek to ensure that important questions raised in relation to artificial intelligence systems are, as appropriate, duly considered through public discussion and multistakeholder consultation in the light of social, economic, legal, ethical, environmental and other relevant implications.

Article 20 – Digital literacy and skills

Each Party shall encourage and promote adequate digital literacy and digital skills for all segments of the population, including specific expert skills for those responsible for the identification, assessment, prevention and mitigation of risks posed by artificial intelligence systems.

Article 21 – Safeguard for existing human rights

Nothing in this Convention shall be construed as limiting, derogating from or otherwise affecting the human rights or other related legal rights

and obligations which may be guaranteed under the relevant laws of a Party or any other relevant international agreement to which it is party.

Article 22 – Wider protection

None of the provisions of this Convention shall be interpreted as limiting or otherwise affecting the possibility for a Party to grant a wider measure of protection than is stipulated in this Convention.

19.2.7 Chapter VII – Follow-up mechanism and co-operation

Article 23 – Conference of the Parties

1. The Conference of the Parties shall be composed of representatives of the Parties to this Convention.

2. The Parties shall consult periodically with a view to:

a. facilitating the effective application and implementation of this Convention, including the identification of any problems and the effects of any reservation made in pursuance of Article 34, paragraph 1, or any declaration made under this Convention;

b. considering the possible supplementation to or amendment of this Convention;

c. considering matters and making specific recommendations concerning the interpretation and application of this Convention;

d. facilitating the exchange of information on significant legal, policy or technological developments of relevance, including in pursuit of the objectives defined in Article 25, for the implementation of this Convention;

e. facilitating, where necessary, the friendly settlement of disputes related to the application of this Convention; and

f. facilitating co-operation with relevant stakeholders concerning pertinent aspects of the implementation of this Convention, including through public hearings where appropriate.

3. The Conference of the Parties shall be convened by the Secretary General of the Council of Europe whenever necessary and, in any case, when a majority of the Parties or the Committee of Ministers requests its convocation.

4. The Conference of the Parties shall adopt its own rules of procedure by consensus within twelve months of the entry into force of this Convention.

5. The Parties shall be assisted by the Secretariat of the Council of Europe in carrying out their functions pursuant to this article.

6. The Conference of the Parties may propose to the Committee of Ministers appropriate ways to engage relevant expertise in support of the effective implementation of this Convention.

7. Any Party which is not a member of the Council of Europe shall contribute to the funding of the activities of the Conference of the Parties. The contribution of a non-member of the Council of Europe shall be established jointly by the Committee of Ministers and that non-member.

8. The Conference of the Parties may decide to restrict the participation in its work of a Party that has ceased to be a member of the Council of Europe under Article 8 of the Statute of the Council of Europe (ETS No. 1) for a serious violation of Article 3 of the Statute. Similarly, measures can be taken in respect of any Party that is not a member State of the Council of Europe by a decision of the Committee of Ministers to cease its relations with that State on grounds similar to those mentioned in Article 3 of the Statute.

Article 24 – Reporting obligation

1. Each Party shall provide a report to the Conference of the Parties within the first two years after becoming a Party, and then periodically thereafter with details of the activities undertaken to give effect to Article 3, paragraph 1, sub-paragraphs a and b.

2. The Conference of the Parties shall determine the format and the process for the report in accordance with its rules of procedure.

Article 25 – International co-operation

1. The Parties shall co-operate in the realisation of the purpose of this Convention. Parties are further encouraged, as appropriate, to assist States that are not Parties to this Convention in acting consistently with the terms of this Convention and becoming a Party to it.

2. The Parties shall, as appropriate, exchange relevant and useful information between themselves concerning aspects related to artificial intelligence which may have significant positive or negative effects on the enjoyment of human rights, the functioning of democracy and the observance of the rule of law, including risks and effects that have arisen in research contexts and in relation to the private sector. Parties are encouraged to involve, as appropriate, relevant stakeholders and States that are not Parties to this Convention in such exchanges of information.

3. The Parties are encouraged to strengthen co-operation, including with relevant stakeholders where appropriate, to prevent and mitigate risks and adverse impacts on human rights, democracy and the rule of law in the context of activities within the lifecycle of artificial intelligence systems.

Article 26 – Effective oversight mechanisms

1. Each Party shall establish or designate one or more effective mechanisms to oversee compliance with the obligations in this Convention.

2. Each Party shall ensure that such mechanisms exercise their duties independently and impartially and that they have the necessary powers, expertise and resources to effectively fulfil their tasks of overseeing compliance with the obligations in this Convention, as given effect by the Parties.

3. If a Party has provided for more than one such mechanism, it shall take measures, where practicable, to facilitate effective cooperation among them.

4. If a Party has provided for mechanisms different from existing human rights structures, it shall take measures, where practicable, to promote effective cooperation between the mechanisms referred to in paragraph 1 and those existing domestic human rights structures.

19.2.8 Chapter VIII – Final clauses

Article 27 – Effects of the Convention

1. If two or more Parties have already concluded an agreement or treaty on the matters dealt with in this Convention, or have otherwise established relations on such matters, they shall also be entitled to apply that agreement or treaty or to regulate those relations accordingly, so long as they do so in a manner which is not inconsistent with the object and purpose of this Convention.

2. Parties which are members of the European Union shall, in their mutual relations, apply European Union rules governing the matters within the scope of this Convention without prejudice to the object and purpose of this Convention and without prejudice to its full application with other Parties. The same applies to other Parties to the extent that they are bound by such rules.

Article 28 – Amendments

1. Amendments to this Convention may be proposed by any Party, the Committee of Ministers of the Council of Europe or the Conference of the Parties.

2. Any proposal for amendment shall be communicated by the Secretary General of the Council of Europe to the Parties.

3. Any amendment proposed by a Party, or the Committee of Ministers, shall be communicated to the Conference of the Parties, which shall

submit to the Committee of Ministers its opinion on the proposed amendment.

4. The Committee of Ministers shall consider the proposed amendment and the opinion submitted by the Conference of the Parties and may approve the amendment.

5. The text of any amendment approved by the Committee of Ministers in accordance with paragraph 4 shall be forwarded to the Parties for acceptance.

6. Any amendment approved in accordance with paragraph 4 shall come into force on the thirtieth day after all Parties have informed the Secretary General of their acceptance thereof.

Article 29 – Dispute settlement

In the event of a dispute between Parties as to the interpretation or application of this Convention, these Parties shall seek a settlement of the dispute through negotiation or any other peaceful means of their choice, including through the Conference of the Parties, as provided for in Article 23, paragraph 2, sub-paragraph e.

Article 30 – Signature and entry into force

1. This Convention shall be open for signature by the member States of the Council of Europe, the non-member States which have participated in its elaboration and the European Union.

2. This Convention is subject to ratification, acceptance or approval. Instruments of ratification, acceptance or approval shall be deposited with the Secretary General of the Council of Europe.

3. This Convention shall enter into force on the first day of the month following the expiration of a period of three months after the date on which five signatories, including at least three member States of the Council of Europe, have expressed their consent to be bound by this Convention in accordance with paragraph 2.

4. In respect of any signatory which subsequently expresses its consent to be bound by it, this Convention shall enter into force on the first day of the month following the expiration of a period of three months after the date of the deposit of its instrument of ratification, acceptance or approval.

Article 31 – Accession

1. After the entry into force of this Convention, the Committee of Ministers of the Council of Europe may, after consulting the Parties to this Convention and obtaining their unanimous consent, invite any non-member State of the Council of Europe which has not participated in the elaboration of this Convention to accede to this Convention by a decision taken by the majority provided for in Article 20.d of the Statute of the Council of Europe, and by unanimous vote of the representatives of the Parties entitled to sit on the Committee of Ministers.

2. In respect of any acceding State, this Convention shall enter into force on the first day of the month following the expiration of a period of three months after the date of deposit of the instrument of accession with the Secretary General of the Council of Europe.

Article 32 – Territorial application

1. Any State or the European Union may, at the time of signature or when depositing its instrument of ratification, acceptance, approval or accession, specify the territory or territories to which this Convention shall apply.

2. Any Party may, at a later date, by a declaration addressed to the Secretary General of the Council of Europe, extend the application of this Convention to any other territory specified in the declaration. In respect of such territory, this Convention shall enter into force on the first day of the month following the expiration of a period of three months after the date of receipt of the declaration by the Secretary General.

3. Any declaration made under the two preceding paragraphs may, in respect of any territory specified in said declaration, be withdrawn by a notification addressed to the Secretary General of the Council of Europe. The withdrawal shall become effective on the first day of the month following the expiration of a period of three months after the date of receipt of such notification by the Secretary General.

Article 33 – Federal clause

1. A federal State may reserve the right to assume obligations under this Convention consistent with its fundamental principles governing the relationship between its central government and constituent states or other similar territorial entities, provided that this Convention shall apply to the central government of the federal State.

2. With regard to the provisions of this Convention, the application of, which come under the jurisdiction of constituent states or other similar territorial entities that are not obliged by the constitutional system of the federation to take legislative measures, the federal government shall inform the competent authorities of such states of the said provisions with its favourable opinion, and encourage them to take appropriate action to give them effect.

Article 34 – Reservations

1. By a written notification addressed to the Secretary General of the Council of Europe, any State may, at the time of signature or when depositing its instrument of ratification, acceptance, approval or accession, declare that it avails itself of the reservation provided for in Article 33, paragraph 1.

2. No other reservation may be made in respect of this Convention.

Article 35 – Denunciation

1. Any Party may, at any time, denounce this Convention by means of a notification addressed to the Secretary General of the Council of Europe.

2. Such denunciation shall become effective on the first day of the month following the expiration of a period of three months after the date of receipt of the notification by the Secretary General.

Article 36 – Notification

The Secretary General of the Council of Europe shall notify the member States of the Council of Europe, the non-member States which have participated in the elaboration of this Convention, the European Union, any signatory, any contracting State, any Party and any other State which has been invited to accede to this Convention, of:

a. any signature;

b. the deposit of any instrument of ratification, acceptance, approval or accession;

c. any date of entry into force of this Convention, in accordance with Article 30, paragraphs 3 and 4, and Article 31, paragraph 2;

d. any amendment adopted in accordance with Article 28 and the date on which such an amendment enters into force;

e. any declaration made in pursuance of Article 3, paragraph 1, subparagraph b;

f. any reservation and withdrawal of a reservation made in pursuance of Article 34;

g. any denunciation made in pursuance of Article 35;

h. any other act, declaration, notification or communication relating to this Convention.

In witness whereof the undersigned, being duly authorised thereto, have signed this Convention.

Done in Vilnius, this 5th day of September 2024, in English and in French, both texts being equally authentic, in a single copy which shall be deposited in the archives of the Council of Europe. The Secretary General of the Council of Europe shall transmit certified copies to each member State of the Council of Europe, to the non-member States which have

participated in the elaboration of this Convention, to the European Union and to any State invited to accede to this Convention.

20

AFRICAN UNION: AFRICAN DIGITAL COMPACT

20.1 Introduction

The "African Digital Compact" (ADC) of the African Union is a 100 pages document, published in July 2024 by the African Union Commission[429]. It is part of the African Union Agenda 2063. The ADC underlines the importance of multi-stakeholder global digital cooperation as well as between African countries. The following excerpt is the Executive Summary (pp 6-9) and the chapter on Development of Artificial Intelligence in the African Context (pp.34-37).

The Editors

20.2 African Digital Compact. Executive summary

"If you want to go fast, to alone. If you want to go far, go together." (African Proverb). This proverb highlights the essence of collaboration and collective action, aligning with the ADC's focus on multi-stakeholder digital cooperation. It underscores the importance of unified efforts across African nations and various sectors to harness the power of

[429] African Union Commission, *African Digital Compact,* Addis Ababa, July 2024. Free download https://au.int.

digital technologies for sustainable development, inclusivity, and economic growth, resonating with the goals of both the ADC and Agenda 2063. This proverb captures the spirit of cooperation and long-term commitment needed to realize a digitally empowered and inclusive Africa.

Vision and purpose

The African Union Digital Compact (ADC) is a forward-looking initiative aimed at harnessing the transformative potential of digital technologies to foster sustainable development, economic growth, and societal well-being throughout Africa. It is Africa's contribution to the development and negotiations of the United Nations Global Digital Compact in preparation of the Summit of the Future, the African Union's Agenda 2063, the Digital Transformation Strategy for Africa (2020-2030) and its associated policies, frameworks and strategies such as the AU Interoperability Framework for Digital ID, AU Data Policy Framework, the AU Strategy on Enabling Policy and Regulatory Environment for Africa's Digital Single Market, Child Online Safety and Empowerment Policy, and the Malabo Convention on Cyber Security and Personal Data Protection. The Compact seeks to unify Africa's digital transformation efforts, ensuring they align with global standards while addressing the continent's unique challenges and opportunities. The ADC vision is to realize an inclusive, resilient, and sustainable digital Africa that leverages ICTs for the continent's socio-economic development and integration, fostering innovation and digital rights for all its citizens, in alignment with Agenda 2063's aspirations for a prosperous Africa based on inclusive growth and sustainable development. The main purpose of ADC is to foster stakeholder digital cooperation within Africa and beyond, aligning with global standards and practices to bridge digital divides, enhance digital literacy, and ensure that digital technologies serve as a catalyst for achieving the SDGs and Agenda 2063 goals. The compact aims to promote digital inclusivity, protect digital rights, and ensure a safe and

secure digital environment for all Africans, and position Africa as a proactive contributor to the global digital economy. This document distils the core aspects and strategic imperatives outlined in the comprehensive analysis and recommendations provided in the context of:

1. The ongoing development of the Global Digital Compact
2. The African Union's Agenda 2063
3. The Digital Transformation Strategy for Africa (2020-2030)
4. The African Union (Malabo) Convention on Cyber Security and Personal Data Protection
5. The Digital Sectorial Strategies for agriculture, education, and health
6. The AU Interoperability Framework for Digital ID
7. The AU Data Policy Framework,
8. The AU Strategy on Enabling Policy and Regulatory Environment for Africa's Digital Single Market
9. The Africa Continental Free Trade Agreement (AfCFTA)
10. The unique opportunities and challenges presented by Africa's digital landscape, all underpinned by guiding principles tailored to the African context.

Strategic imperatives

The Compact outlines several strategic imperatives:

1. Universal connectivity and universal access: Aiming to connect all individuals and institutions, safely and securely to the Internet, and empowering them to fully utilize digital technologies and services, regardless of their gender, age, education, geographical local, socioeconomic status, or any other divides.

2. Capacity building to empower Africa's engagement in the digital economy: Accelerating knowledge transfer and adoption, and on enhancing digital literacy and skills development, to empower African's engagement in, contribution to, and benefit from the digital economy,

locally and globally. A special focus is required to reap the benefits of the increasing job opportunities in digital technologies and related services, while minimizing the potential impact on the traditional employment.

3. Leveraging digital technologies for green, resilient, and inclusive sustainable development: Mainstreaming digitalization in critical sectors for green and inclusive sustainable development

4. Data protection and internet integrity: Safeguarding privacy, enhancing data governance, and avoiding Internet fragmentation.

5. Cybersecurity and the protection of the ICT-enabled critical infrastructures: Establish and implement frameworks for cybersecurity and for the protection of the ICT-enabled critical infrastructures.

6. Safety and stability: Effective engagement with international and regional initiatives and contribution towards developing and implementing cyber norms and confidence building measures.

7. Digital rights and accountability: Applying human rights principles online and ensuring transparency and accountability online.

8. Regulation and innovation: Encouraging the ethical regulation of artificial intelligence and other emerging technologies and supporting innovation and creativity across Africa.

9. Digital governance: Bridging the digital governance gap for African countries to support Africa's Socio-economic development and reap the full potential of the digital economy.

Collaborative framework

The Compact's development involved extensive consultations with stakeholders across Africa, including the public and private sectors, civil society, academia, and regional organizations. This inclusive approach ensured that the Compact reflects a wide range of perspectives and priorities, embodying a collective vision for Africa's digital future.

Objectives and implementation

The Compact emphasizes the need for:

1. Bridging the digital divides by making digital technologies accessible and affordable across the continent.

2. Enhancing digital literacy and skills to prepare Africans for the digital economy.

3. Leveraging Digital Technologies for Green, Resilient, and Inclusive Sustainable Development

4. Developing and implementing robust cybersecurity strategies and policies to protect digital ecosystems.

5. Supporting digital innovation and entrepreneurship to stimulate economic growth.

6. Harmonizing digital and data policies to facilitate cross-border collaboration.

7. Bridging the digital governance gap to accelerate economic growth and maximize safe and peaceful use of ICTs.

To support the implementation of the African Digital Compact, an African Digital Cooperation Forum (ADCF) is being proposed as a *Pioneering Platform for Multi-stakeholder Engagement, Collaboration, and Innovation in the Digital Sphere*. This forum Serves as a Knowledge Broker and an Investment Facilitator for collaborative ADC programs and projects, and a conduit for harmonizing digital initiatives across the continent and beyond.

Moving forward

The African Union Digital Compact calls for unified action among African nations, the private sector, civil society, and international partners. It envisions a digitally empowered Africa where digital technologies drive progress, inclusivity, and prosperity. This executive summary invites all stakeholders to join in realizing this vision, emphasizing that collective effort is essential for a sustainable and inclusive digital

transformation and socio-economic development across Africa to accelerate the attainment of the AU Agenda 2063 goals and aspirations.

20.3 African Digital Compact, pillar eight: Development and adoption of artificial intelligence (AI)

Objective: Artificial Intelligence (AI) is a transformative and revolutionary technology that has a huge potential to stimulate economic growth and support sustainable development by creating new industries, driving innovation, and generating employment opportunities, it impacts every sector ranging from finance, national security, health care, education, criminal justice, transportation, smart cities, private and public life. To effectively Leverage AI Potential there is a need to address the societal, security, ethical and legal challenges associated with AI that may affect Africa society and people.

Vision and strategic objectives: Ensure Africa-centric, development focused, Responsible and Ethical AI that Empowers People and Contributes to the Continent Inclusive Growth, Resilience and Socioeconomic Progress. The main Objectives associated with AI adoption in Africa are:

• Develop and implement robust AI governance, regulations, standards, codes of conduct and best practices to manage AI risks and promote its growth.

• Accelerate the integration of AI in the core sectors notably sectors with high social and economic value, including agriculture, education, health, climate change and natural resource management, and regional peace and security. Accelerate the adoption of AI by the private sector, including small and medium-sized enterprises and create an enabling environment for AI start-up ecosystem focused on solving Africa development problems.

• Ensure the availability of high-quality and diverse datasets to reflect Africa cultural identity and diversity, underpin AI development and

ensure the availability of AI infrastructure, such as data centres, computing platforms, and IoTs to support the development of local AI-based systems, solutions and products.

• Encourage research and innovation and Promote inclusivity and diversity in AI skills and AI talent to prepare Africa's workforce of tomorrow.

• Promote regional and international cooperation to maximise the benefits of AI, minimise its risks, share capabilities and resources and ensure equal opportunities for African people and communities.

Key policy interventions

1. Strengthen AI governance

1.1 Develop National AI Strategies and advance a multi-tiered governance and institutional approach.

1.2 Promote agile, forward-looking and risk-based regulations at national and regional levels that promote accountability and transparency in the design and deployment of AI Systems.

1.3 Establish oversight bodies at national and regional levels to monitor AI developments and ensure compliance with ethical standards and Africa values.

2. Optimizing AI benefits for socioeconomic development

2.1 Promote the use of AI in key sectors to increase social and economic benefits such as public sector, education, health, agriculture and facilitate exchanges of knowledge and experience on use cases among AU Member States.

2.2 Raise awareness among Member States of the potential benefits of adopting AI to mitigate, adapt to, and build resilience against the risks of climate change.

3. Building Africa's AI capabilities

3.1 Raise public awareness on the opportunities and challenges stemming from AI.

3.2 Sensitize on the need for developing Africa-centric, diversified and reliable datasets, promote national and regional data pools and data markets to support the development of AI models that respond to Africa needs.

3.3 Support academic and industry-led research initiatives focused on ethical and responsible AI development that addresses Africa's unique challenges and opportunities.

3.4 Develop inclusive and gender responsive policies on AI skills and competencies to leverage AI for socio-economic development.

4. Addressing the risk and threats arising from AI

4.1 Work towards reducing biases and closing the gender, socio-economic, cultural and rural-urban gaps and ensure equity, justice and equal opportunities to all African citizens in the development and adoption of AI Systems.

4.2 Develop Media and Information Literacy Policies to address the AI generated misinformation, disinformation and hate speech to preserve the cohesion of African Societies, people well-being, national economies and democracies.

4.3 Upgrade cybersecurity measures to ensure highest standards of AI safety and security in the lifecycle of AI systems and solutions in Africa.

4.4 Assess the implication of AI on Africa Peace and Security landscape.

5. Foster public and private sectors cooperation and investment in AI

5.1 Forge partnerships between the public and private sectors and increase investment in AI (research, skills, research and innovation ...).

5.2 Promote private sector AI readiness, create an enabling environment and provide guidance and policy actions, especially regarding small and medium enterprises' adoption of AI in the areas where the country has a comparative advantage.

5.3 Promote AI startups as engines of inclusive growth, thus ensuring all incentives are in place to attract and retain talent and enterprises in the region.

6. Promote regional and international cooperation on AI

6.1 Foster intra-Africa cooperation on AI through exchange of AI expertise across the Continent, developing joint research programs and creating a network of centers of Excellence to address Africa needs.

6.2 Leverage AU Strategic Partnerships at the Multilateral Level to include AI as area of cooperation work towards bridging the AI and technological gap between Africa and other regions.

6.3 Actively participate in International discussions on the global governance of AI to ensure African perspectives and interests are preserved.

Implementation and collaboration

1. Multi-stakeholder engagement: Foster a collaborative approach by engaging all relevant stakeholders in the development, deployment, and governance of AI, ensuring a multi-disciplinary perspective that includes ethical, cultural, and socioeconomic considerations.

2. Policy and regulatory frameworks: Encourage Member States to develop and harmonize AI policies and regulations that reflect the principles outlined in the AU Strategy on Artificial Intelligence, ensuring a conducive environment for AI innovation and AI uptake across Africa.

3. Resource allocation: Mobilize resources, both financial and technical, to support research, development, and deployment of AI solutions that are in line with Africa local context and needs.

4. Monitoring and evaluation: Establish mechanisms for monitoring and evaluation of AI adoption by AU Member States towards the development of a consolidate AI ecosystem in Africa. The Ethical, responsible, Africa centric and development-oriented AI represents a critical pillar

within the African Digital Compact, reflecting Africa's commitment to harnessing AI as a force for good to transform its economy and society.

21

CHINA: STRENGTHENING ETHICAL GLOBAL GOVERNANCE OF AI AND THE SHANGHAI DECLARATION

21.1 Strengthening ethical governance of AI

21.1.1 Introduction

The People's Republic of China has published a number of relevant policy documents on Artificial Intelligence[430]. The China Ministry of Foreign Affairs published the following "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)". It was published in 2022 and updated on November 17, 2022.[431] Compared to other countries, China is advanced in guidelines and governance rules for AI, as it is also fast developing AI in many sectors, from industry to education, from infrastructure to military, from agriculture to water management.

The Editors

21.1.2 The position paper

I. As the most representative disruptive technology, artificial intelligence (AI), while providing enormous potential development benefits

[430] See the overview in article 11.2.1 in this book.

[431] https://www.fmprc.gov.cn/eng/zy/wjzc/202405/t20240531_11367525.html.

to human society, has also brought uncertainty that may give rise to multiple global challenges and even fundamental ethical concerns. At the ethical level, there are widespread concerns in the international community that, if left unregulated, the misuse and abuse of AI technologies may undermine human dignity and equality, violate human rights and fundamental freedoms, exacerbate discrimination and prejudice, disrupt existing legal systems, and have far-reaching impacts on government administration, building of national defence, social stability and even global governance.

China is committed to building a community with a shared future for mankind in the domain of AI, advocating a people-cantered approach and the principle of AI for good. China finds it important to enhance the understanding of all countries on AI ethics, and ensure that AI is safe, reliable, controllable, and capable of better empowering global sustainable development and enhancing the common well-being of all mankind. To this end, China calls on all parties to uphold the principles of extensive consultation, joint contribution and shared benefits, and advance international AI ethical governance.

II. In December 2021, China released the Position Paper of the Peoples' Republic of China on Regulating Military Applications of Artificial Intelligence (AI), calling on parties to observe national or regional ethical norms for AI. In this connection, China, based on its own policies and practices and with reference to useful international experience, makes the following points in the aspects of regulation, research and development, utilization and international cooperation.

(1) Regulation

Governments should give priority to ethics, establish and improve rules, norms and accountability mechanisms for AI ethics, clarify responsibilities and power boundaries of AI-related entities, fully respect and

protect the legitimate rights and interests of all groups, and respond to relevant ethical concerns at home and abroad in a timely manner.

Governments should place emphasis on research on fundamental theoretical issues regarding AI ethics and laws, gradually put in place and improve AI ethical norms, laws, regulations, and policies, formulate guidelines on AI ethics, establish review and regulation mechanisms for ethics in science and technology, and strengthen evaluation and management capacity for AI security.

Governments should bear in mind worst-case scenarios and enhance risk awareness, better identify potential ethical risks that AI technologies may entail, gradually establish an effective early warning mechanism, apply agile governance and tiered and categorized management, and continuously improve risk management, control and settlement capacity.

Governments should, in light of their own stage of AI development as well as social and cultural characteristics, and in accordance with the new features of scientific and technological innovation, gradually establish AI ethical systems suited to their national conditions, and improve AI ethical governance mechanisms featuring wide participation and cooperative governance.

(2) Research and development (R&D)

Governments should require R&D entities to strengthen self-discipline, voluntarily incorporate ethics into the whole process of AI R&D, avoid premature use of technologies that may cause serious consequences, and ensure that AI is always under the control of humans.

Governments should require R&D entities to strive for algorithm security and controllability throughout the AI R&D process, improve transparency, explainability and reliability in the stages of algorithm design, implementation, application, etc., and gradually make AI verifiable, regulatable, traceable, predictable and trustworthy.

Governments should require R&D entities to strive for better data quality during AI R&D, strictly abide by data security regulations, ethics and relevant laws and standards in the phases of data collection, storage, utilization, etc., and improve the completeness, timeliness, consistency, normalization and accuracy of data.

Governments should require R&D entities to strengthen ethics review of data collection and algorithm development, fully consider diversified demands, avoid potential data and algorithms bias, and strive to achieve the universality, fairness and non-discrimination of AI systems.

(3) Utilization

Governments should prohibit using AI technologies and relevant applications which run counter to laws, regulations, ethics and standards, strengthen quality monitoring and evaluations on the use of AI products and services, and formulate emergency mechanisms and compensation measures.

Governments should strengthen pre-use study and evaluations of AI products and services, and promote institutionalized training on AI ethics. Relevant personnel should fully understand the functions, features, limitations and potential risks and consequences of AI technologies, and acquire necessary professional expertise and skills.

Governments should safeguard individual privacy and data security of Al products and services, strictly follow international or regional norms for handling of personal information, improve the mechanism for revoking personal data authorization, and oppose illegal collection and utilization of personal information.

Governments should attach importance to public education on AI ethics, guarantee the public's rights of knowledge and meaningful participation, give full play to the roles of scientific and technological communities, guide all sectors of society to comply with AI ethical rules and norms, and raise AI ethical awareness.

(4) International cooperation

Governments should encourage transnational, interdisciplinary, and cross-cultural exchanges and cooperation, ensure that the benefits of AI technologies are shared by all countries, promote joint participation of countries in international discussions and rules-making on major issues regarding AI ethics, and oppose the building of exclusive groups and malicious obstruction of other countries' technological development.

Governments should strengthen the regulation of AI ethics for international cooperative research activities. Relevant science and technology activities should comply with the requirements of AI ethics management in the countries where the cooperating parties are located, and pass the AI ethics review accordingly.

China calls on the international community to reach international agreement on the issue of AI ethics on the basis of wide participation, and work to formulate widely accepted international AI governance framework, standards and norms while fully respecting the principles and practices of different countries' AI governance.

21.2 Shanghai Declaration on Global AI Governance

21.2.1 Introduction

The People's Republic of China published a number of relevant policy documents on AI[432]. The 2024 "World AI Conference and High-Level Meeting on Global AI Governance" was held in July 2024 in Shanghai. The China Ministry of Foreign Affairs published the following "Shanghai Declaration on Global AI Governance" on July 04, 2024. It is the full text of the short declaration.

[432] See the overview in article 11.2.1 in this book.

21.2.2 Shanghai Declaration

We are fully aware of the far-reaching impact of artificial intelligence (AI) on the world and its great potential, and acknowledge that AI is leading a scientific and technological revolution and profoundly affecting the way people work and live. With the rapid development of AI technologies, we are also facing unprecedented challenges, especially in terms of safety and ethics.

We underline the need to promote the development and application of AI technologies while ensuring safety, reliability, controllability and fairness in the process, and encourage leveraging AI technologies to empower the development of human society. We believe that only through global cooperation and a collective effort can we realize the full potential of AI for the greater well-being of humanity.

(1) Promoting AI development

We agree to actively promote research and development to unleash the potential of AI in various fields such as healthcare, education, transportation, agriculture, industry, culture and ecology. We will encourage innovative thinking, support interdisciplinary research collaboration, and jointly promote breakthroughs of AI technologies and AI for good. We will closely watch and mitigate the impact of AI on employment, and guide and promote the improvement of the quality and efficiency of AI-enabled human work.

We advocate the spirit of openness and shared benefit, and will promote exchanges and cooperation on global AI research resources. We will establish cooperation platforms to facilitate technology transfer and commercialization, promote fair distribution of AI infrastructure, avoid technical barriers, and jointly strengthen global AI development.

We agree to safeguard high-quality data development with high-level data security, promote the free and orderly flow of data in accordance with the law, oppose discriminatory and exclusive data training, and

collaborate in the development of high-quality datasets, so as to better nourish AI development.

We will establish cooperation mechanisms to vigorously promote AI empowerment across industries, starting with accelerating smart application in such fields as manufacturing, logistics and mining, and simultaneously promoting the sharing of relevant technologies and standards.

We are committed to cultivating more AI professionals, strengthening education, training and personnel exchanges and cooperation, and improving AI literacy and skills around the world.

We call upon all countries to uphold a people-centred approach and adhere to the principle of AI for good, and ensure equal rights, equal opportunities and equal rules for all countries in developing and using AI technologies without any form of discrimination.

We respect the right of all countries to independent development, encourage all countries to formulate AI strategies, policies and laws and regulations based on their own national conditions, and call for abiding by the laws and regulations of countries receiving the goods and services, observing applicable international law, and respecting their economic and social systems, religious and cultural traditions and values in carrying out international cooperation on AI technologies, products and applications.

(2) Maintaining AI safety

We attach great importance to AI safety, especially to data security and privacy protection. We agree to promote the formulation of data protection rules, strengthen the interoperability of data and information pro(t)ction policies of different countries, and ensure the protection and lawful use of personal information.

We recognize the need to strengthen regulation, and develop reliable AI technologies that can be reviewed, monitored and traced. Bearing in mind the evolving nature of AI, we will use AI technologies to prevent

AI risks and enhance the technological capacity for AI governance, on the basis of human decision-making and supervision. We encourage countries, in light of their national conditions, to formulate laws and norms, and establish a testing and assessment system based on AI risk levels and a sci-tech ethical review system. On this basis, we encourage the formulation of more timely and agile self-discipline norms for the industry.

We resolve to strengthen AI-related cybersecurity, enhance the security and reliability of systems and applications, and prevent hacking and malware applications. We decide to jointly combat the use of AI to manipulate public opinion and fabricate and disseminate disinformation on the premise of respecting and applying international and domestic legal frameworks.

We will work together to prevent terrorists, extremist forces, and transnational organized criminal groups from using AI technologies for illegal activities, and jointly combat the theft, tampering, leaking and illegal collection and use of personal information.

We agree to promote the formulation and adoption of ethical guidelines and norms for AI with broad international consensus, guide the healthy development of AI technologies, and prevent their misuse, abuse or malicious use.

(3) Developing the AI governance system

We advocate establishing an AI governance mechanism of a global scope, support the role of the United Nations as the main channel, welcome the strengthening of North-South and South-South cooperation, and call for increasing the representation and voice of developing countries. We encourage various actors including international organizations, enterprises, research institutes, social organizations, and individuals to actively play their due roles in the development and implementation of the AI governance system.

We agree to strengthen cooperation with international organizations and professional institutes to share policies and practices of AI testing, assessment, certification and regulation to ensure the safety, controllability and reliability of AI technologies.

We agree to strengthen the regulatory and accountability mechanisms for AI to ensure compliance and accountability in the use of AI technologies.

(4) Strengthening public participation and improving literacy

We agree to establish mechanisms for diverse participation, including public consultation, social surveys, etc., to include the public in decision-making on AI.

We will increase the public's knowledge and understanding of AI and raise public awareness about AI safety. We will carry out communication activities to popularize AI knowledge and enhance digital literacy and safety awareness among the public.

(5) Improving quality of life and increasing social well-being

We will actively promote the application of AI in the field of sustainable development, including industrial innovation, environmental protection, resource utilization, energy management, and biodiversity promotion. We encourage innovative thinking in exploring the potential of AI technologies in contributing to the resolution of global issues.

We are committed to using AI to improve social well-being, especially in such fields as healthcare, education, and elderly care.

We are fully aware that the implementation of this declaration requires our joint efforts. We look forward to positive responses from governments, sci-tech communities, industrial communities and other stakeholders around the world. Together, let us promote the healthy development of AI, ensure AI safety, and empower the common future of mankind with AI.

22

UNITED STATES: EXECUTIVE ORDER ON THE SAFE, SECURE AND TRUSTWORTHY DEVELOPMENT AND USE OF ARTIFICIAL INTELLIGENCE

22.1 Introduction

US President Joe Biden published on 31 October 2023 the "Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence"[433]. The detailed order with precise timelines for implementation includes eight principles, addressed to all sectors of society. The fact to decide it as presidential order and its content shows, that the US president expressed the urgency of the matter as AI develops very fast. The order is primarily addressed to the actors in the United States, but also calls for multilateral, international cooperation, which is needed for a "safe, secure and trustworthy" development and use of AI. The order includes ten sections on safety and security, innovation and competition,

[433] Accessible at: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

supporting workers, advancing equity and civil rights, protecting consumers, patients, students, protecting privacy

The Editors

22.2 Executive order on the safe, secure and trustworthy development and use of artificial intelligence

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

22.2.1 Section 1. purpose

Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavour demands a society-wide effort that includes government, the private sector, academia, and civil society.

My Administration places the highest urgency on governing the development and use of AI safely and responsibly, and is therefore advancing a coordinated, Federal Government-wide approach to doing so. The rapid speed at which AI capabilities are advancing compels the United States to lead in this moment for the sake of our security, economy, and society.

In the end, AI reflects the principles of the people who build it, the people who use it, and the data upon which it is built. I firmly believe that the power of our ideals; the foundations of our society; and the creativity, diversity, and decency of our people are the reasons that America

thrived in past eras of rapid change. They are the reasons we will succeed again in this moment. We are more than capable of harnessing AI for justice, security, and opportunity for all.

22.2.2 Section 2. Policy and principles

It is the policy of my Administration to advance and govern the development and use of AI in accordance with eight guiding principles and priorities. When undertaking the actions set forth in this order, executive departments and agencies (agencies) shall, as appropriate and consistent with applicable law, adhere to these principles, while, as feasible, taking into account the views of other agencies, industry, members of academia, civil society, labor unions, international allies and partners, and other relevant organizations:

(a) Artificial Intelligence must be safe and secure. Meeting this goal requires robust, reliable, repeatable, and standardized evaluations of AI systems, as well as policies, institutions, and, as appropriate, other mechanisms to test, understand, and mitigate risks from these systems before they are put to use. It also requires addressing AI systems' most pressing security risks — including with respect to biotechnology, cybersecurity, critical infrastructure, and other national security dangers — while navigating AI's opacity and complexity. Testing and evaluations, including post-deployment performance monitoring, will help ensure that AI systems function as intended, are resilient against misuse or dangerous modifications, are ethically developed and operated in a secure manner, and are compliant with applicable Federal laws and policies. Finally, my Administration will help develop effective labelling and content provenance mechanisms, so that Americans are able to determine when content is generated using AI and when it is not. These actions will provide a vital foundation for an approach that addresses AI's risks without unduly reducing its benefits.

(b) Promoting responsible innovation, competition, and collaboration will allow the United States to lead in AI and unlock the technology's potential to solve some of society's most difficult challenges. This effort requires investments in AI-related education, training, development, research, and capacity, while simultaneously tackling novel intellectual property (IP) questions and other problems to protect inventors and creators. Across the Federal Government, my Administration will support programs to provide Americans the skills they need for the age of AI and attract the world's AI talent to our shores — not just to study, but to stay — so that the companies and technologies of the future are made in America. The Federal Government will promote a fair, open, and competitive ecosystem and marketplace for AI and related technologies so that small developers and entrepreneurs can continue to drive innovation. Doing so requires stopping unlawful collusion and addressing risks from dominant firms' use of key assets such as semiconductors, computing power, cloud storage, and data to disadvantage competitors, and it requires supporting a marketplace that harnesses the benefits of AI to provide new opportunities for small businesses, workers, and entrepreneurs.

(c) The responsible development and use of AI require a commitment to supporting American workers. As AI creates new jobs and industries, all workers need a seat at the table, including through collective bargaining, to ensure that they benefit from these opportunities. My Administration will seek to adapt job training and education to support a diverse workforce and help provide access to opportunities that AI creates. In the workplace itself, AI should not be deployed in ways that undermine rights, worsen job quality, encourage undue worker surveillance, lessen market competition, introduce new health and safety risks, or cause harmful labor-force disruptions. The critical next steps in AI development should be built on the views of workers, labor unions, educators, and

employers to support responsible uses of AI that improve workers' lives, positively augment human work, and help all people safely enjoy the gains and opportunities from technological innovation.

(d) Artificial Intelligence policies must be consistent with my Administration's dedication to advancing equity and civil rights. My Administration cannot — and will not — tolerate the use of AI to disadvantage those who are already too often denied equal opportunity and justice. From hiring to housing to healthcare, we have seen what happens when AI use deepens discrimination and bias, rather than improving quality of life. Artificial Intelligence systems deployed irresponsibly have reproduced and intensified existing inequities, caused new types of harmful discrimination, and exacerbated online and physical harms. My Administration will build on the important steps that have already been taken — such as issuing the Blueprint for an AI Bill of Rights, the AI Risk Management Framework, and Executive Order 14091 of February 16, 2023 (Further Advancing Racial Equity and Support for Underserved Communities Through the Federal Government) — in seeking to ensure that AI complies with all Federal laws and to promote robust technical evaluations, careful oversight, engagement with affected communities, and rigorous regulation. It is necessary to hold those developing and deploying AI accountable to standards that protect against unlawful discrimination and abuse, including in the justice system and the Federal Government. Only then can Americans trust AI to advance civil rights, civil liberties, equity, and justice for all.

(e) The interests of Americans who increasingly use, interact with, or purchase AI and AI-enabled products in their daily lives must be protected. Use of new technologies, such as AI, does not excuse organizations from their legal obligations, and hard-won consumer protections are more important than ever in moments of technological change. The Federal Government will enforce existing consumer protection laws

and principles and enact appropriate safeguards against fraud, unintended bias, discrimination, infringements on privacy, and other harms from AI. Such protections are especially important in critical fields like healthcare, financial services, education, housing, law, and transportation, where mistakes by or misuse of AI could harm patients, cost consumers or small businesses, or jeopardize safety or rights. At the same time, my Administration will promote responsible uses of AI that protect consumers, raise the quality of goods and services, lower their prices, or expand selection and availability.

(f) Americans' privacy and civil liberties must be protected as AI continues advancing. Artificial Intelligence is making it easier to extract, re-identify, link, infer, and act on sensitive information about people's identities, locations, habits, and desires. Artificial Intelligence's capabilities in these areas can increase the risk that personal data could be exploited and exposed. To combat this risk, the Federal Government will ensure that the collection, use, and retention of data is lawful, is secure, and mitigates privacy and confidentiality risks. Agencies shall use available policy and technical tools, including privacy-enhancing technologies (PETs) where appropriate, to protect privacy and to combat the broader legal and societal risks — including the chilling of First Amendment rights — that result from the improper collection and use of people's data.

(g) It is important to manage the risks from the Federal Government's own use of AI and increase its internal capacity to regulate, govern, and support responsible use of AI to deliver better results for Americans. These efforts start with people, our Nation's greatest asset. My Administration will take steps to attract, retain, and develop public service-oriented AI professionals, including from underserved communities, across disciplines — including technology, policy, managerial, procurement, regulatory, ethical, governance, and legal fields — and ease AI professionals' path into the Federal Government to help harness and

govern AI. The Federal Government will work to ensure that all members of its workforce receive adequate training to understand the benefits, risks, and limitations of AI for their job functions, and to modernize Federal Government information technology infrastructure, remove bureaucratic obstacles, and ensure that safe and rights-respecting AI is adopted, deployed, and used.

(h) The Federal Government should lead the way to global societal, economic, and technological progress, as the United States has in previous eras of disruptive innovation and change. This leadership is not measured solely by the technological advancements our country makes. Effective leadership also means pioneering those systems and safeguards needed to deploy technology responsibly — and building and promoting those safeguards with the rest of the world. My Administration will engage with international allies and partners in developing a framework to manage AI's risks, unlock AI's potential for good, and promote common approaches to shared challenges. The Federal Government will seek to promote responsible AI safety and security principles and actions with other nations, including our competitors, while leading key global conversations and collaborations to ensure that AI benefits the whole world, rather than exacerbating inequities, threatening human rights, and causing other harms.

22.2.3 Section 3. Definitions. For purposes of this order

(a) The term "agency" means each agency described in 44 U.S.C. 3502(1), except for the independent regulatory agencies described in 44 U.S.C. 3502(5).

(b) The term "artificial intelligence" or "AI" 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.

(c) The term "AI model" means a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.

(d) The term "AI red-teaming" means a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.

(e) The term "AI system" means any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI.

(f) The term "commercially available information" means any information or data about an individual or group of individuals, including an individual's or group of individuals' device or location, that is made available or obtainable and sold, leased, or licensed to the general public or to governmental or non-governmental entities.

(g) The term "crime forecasting" means the use of analytical techniques to attempt to predict future crimes or crime-related information. It can include machine-generated predictions that use algorithms to analyze large volumes of data, as well as other forecasts that are generated without machines and based on statistics, such as historical crime statistics.

(h) The term "critical and emerging technologies" means those technologies listed in the February 2022 Critical and Emerging Technologies List Update issued by the National Science and Technology Council (NSTC), as amended by subsequent updates to the list issued by the NSTC.

(i) The term "critical infrastructure" has the meaning set forth in section 1016(e) of the USA PATRIOT Act of 2001, 42 U.S.C. 5195c(e).

(j) The term "differential-privacy guarantee" means protections that allow information about a group to be shared while provably limiting the improper access, use, or disclosure of personal information about particular entities.

(k) The term "dual-use foundation model" means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:

(i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;

(ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or

(iii) permitting the evasion of human control or oversight through means of deception or obfuscation.

Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.

(l) The term "Federal law enforcement agency" has the meaning set forth in section 21(a) of Executive Order 14074 of May 25, 2022 (Advancing Effective, Accountable Policing and Criminal Justice Practices To Enhance Public Trust and Public Safety).

(m) The term "floating-point operation" means any mathematical operation or assignment involving floating-point numbers, which are a subset of the real numbers typically represented on computers by an integer of fixed precision scaled by an integer exponent of a fixed base.

(n) The term "foreign person" has the meaning set forth in section 5(c) of Executive Order 13984 of January 19, 2021 (Taking Additional Steps to Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities).

(o) The terms "foreign reseller" and "foreign reseller of United States Infrastructure as a Service Products" mean a foreign person who has established an Infrastructure as a Service Account to provide Infrastructure as a Service Products subsequently, in whole or in part, to a third party.

(p) The term "generative AI" means the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.

(q) The terms "Infrastructure as a Service Product," "United States Infrastructure as a Service Product," "United States Infrastructure as a Service Provider," and "Infrastructure as a Service Account" each have the respective meanings given to those terms in section 5 of Executive Order 13984.

(r) The term "integer operation" means any mathematical operation or assignment involving only integers, or whole numbers expressed without a decimal point.

(s) The term "Intelligence Community" has the meaning given to that term in section 3.5(h) of Executive Order 12333 of December 4, 1981 (United States Intelligence Activities), as amended.

(t) The term "machine learning" means a set of techniques that can be used to train AI algorithms to improve performance at a task based on data.

(u) The term "model weight" means a numerical parameter within an AI model that helps determine the model's outputs in response to inputs.

(v) The term "national security system" has the meaning set forth in 44 U.S.C. 3552(b)(6).

(w) The term "omics" means biomolecules, including nucleic acids, proteins, and metabolites, that make up a cell or cellular system.

(x) The term "Open RAN" means the Open Radio Access Network approach to telecommunications-network standardization adopted by the O-RAN Alliance, Third Generation Partnership Project, or any similar set of published open standards for multi-vendor network equipment interoperability.

(y) The term "personally identifiable information" has the meaning set forth in Office of Management and Budget (OMB) Circular No. A-130.

(z) The term "privacy-enhancing technology" means any software or hardware solution, technical process, technique, or other technological means of mitigating privacy risks arising from data processing, including by enhancing predictability, manageability, disassociability, storage, security, and confidentiality. These technological means may include secure multiparty computation, homomorphic encryption, zero-knowledge proofs, federated learning, secure enclaves, differential privacy, and synthetic-data-generation tools. This is also sometimes referred to as "privacy-preserving technology."

(aa) The term "privacy impact assessment" has the meaning set forth in OMB Circular No. A-130.

(bb) The term "Sector Risk Management Agency" has the meaning set forth in 6 U.S.C. 650(23).

(cc) The term "self-healing network" means a telecommunications network that automatically diagnoses and addresses network issues to permit self-restoration.

(dd) The term "synthetic biology" means a field of science that involves redesigning organisms, or the biomolecules of organisms, at the genetic level to give them new characteristics. Synthetic nucleic acids are a type of biomolecule redesigned through synthetic-biology methods.

(ee) The term "synthetic content" means information, such as images, videos, audio clips, and text, that has been significantly modified or generated by algorithms, including by AI.

(ff) The term "testbed" means a facility or mechanism equipped for conducting rigorous, transparent, and replicable testing of tools and technologies, including AI and PETs, to help evaluate the functionality, usability, and performance of those tools or technologies.

(gg) The term "watermarking" means the act of embedding information, which is typically difficult to remove, into outputs created by AI — including into outputs such as photos, videos, audio clips, or text — for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.

22.2.4 Section 4. Ensuring the safety and security of AI technology

4.1. Developing guidelines, standards, and best practices for AI safety and security.

(a) Within 270 days of the date of this order, to help ensure the development of safe, secure, and trustworthy AI systems, the Secretary of Commerce, acting through the Director of the National Institute of Standards and Technology (NIST), in coordination with the Secretary of Energy, the Secretary of Homeland Security, and the heads of other relevant agencies as the Secretary of Commerce may deem appropriate, shall:

(i) Establish guidelines and best practices, with the aim of promoting consensus industry standards, for developing and deploying safe, secure, and trustworthy AI systems, including:

(A) developing a companion resource to the AI Risk Management Framework, NIST AI 100-1, for generative AI;

(B) developing a companion resource to the Secure Software Development Framework to incorporate secure development practices for generative AI and for dual-use foundation models; and

(C) launching an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities through which AI could cause harm, such as in the areas of cybersecurity and biosecurity.

(ii) Establish appropriate guidelines (except for AI used as a component of a national security system), including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems. These efforts shall include:

(A) coordinating or developing guidelines related to assessing and managing the safety, security, and trustworthiness of dual-use foundation models; and

(B) in coordination with the Secretary of Energy and the Director of the National Science Foundation (NSF), developing and helping to ensure the availability of testing environments, such as testbeds, to support the development of safe, secure, and trustworthy AI technologies, as well as to support the design, development, and deployment of associated PETs, consistent with section 9(b) of this order.

(b) Within 270 days of the date of this order, to understand and mitigate AI security risks, the Secretary of Energy, in coordination with the heads of other Sector Risk Management Agencies (SRMAs) as the Secretary of Energy may deem appropriate, shall develop and, to the extent permitted by law and available appropriations, implement a plan for developing the Department of Energy's AI model evaluation tools and AI testbeds. The Secretary shall undertake this work using existing solutions where possible, and shall develop these tools and AI testbeds to be capable of assessing near-term extrapolations of AI systems' capabilities. At a minimum, the Secretary shall develop tools to evaluate AI capabilities to generate outputs that may represent nuclear, non-proliferation, biological, chemical, critical infrastructure, and energy-security threats or hazards. The Secretary shall do this work solely for the purposes of guarding against these threats, and shall also develop model guardrails that reduce such risks. The Secretary shall, as appropriate, consult with private AI laboratories, academia, civil society, and third-party evaluators, and shall use existing solutions.

4.2. Ensuring safe and reliable AI.

(a) Within 90 days of the date of this order, to ensure and verify the continuous availability of safe, reliable, and effective AI in accordance

with the Defense Production Act, as amended, 50 U.S.C. 4501 *et seq*., including for the national defense and the protection of critical infrastructure, the Secretary of Commerce shall require:

(i) Companies developing or demonstrating an intent to develop potential dual-use foundation models to provide the Federal Government, on an ongoing basis, with information, reports, or records regarding the following:

(A) any ongoing or planned activities related to training, developing, or producing dual-use foundation models, including the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats;

(B) the ownership and possession of the model weights of any dual-use foundation models, and the physical and cybersecurity measures taken to protect those model weights; and

(C) the results of any developed dual-use foundation model's performance in relevant AI red-team testing based on guidance developed by NIST pursuant to subsection 4.1(a)(ii) of this section, and a description of any associated measures the company has taken to meet safety objectives, such as mitigations to improve performance on these red-team tests and strengthen overall model security. Prior to the development of guidance on red-team testing standards by NIST pursuant to subsection 4.1(a)(ii) of this section, this description shall include the results of any red-team testing that the company has conducted relating to lowering the barrier to entry for the development, acquisition, and use of biological weapons by non-state actors; the discovery of software vulnerabilities and development of associated exploits; the use of software or tools to influence real or virtual events; the possibility for self-replication or propagation; and associated measures to meet safety objectives; and

(ii) Companies, individuals, or other organizations or entities that acquire, develop, or possess a potential large-scale computing cluster to report any such acquisition, development, or possession, including the existence and location of these clusters and the amount of total computing power available in each cluster.

(b) The Secretary of Commerce, in consultation with the Secretary of State, the Secretary of Defense, the Secretary of Energy, and the Director of National Intelligence, shall define, and thereafter update as needed on a regular basis, the set of technical conditions for models and computing clusters that would be subject to the reporting requirements of subsection 4.2(a) of this section. Until such technical conditions are defined, the Secretary shall require compliance with these reporting requirements for:

(i) any model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point operations, or using primarily biological sequence data and using a quantity of computing power greater than 1023 integer or floating-point operations; and

(ii) any computing cluster that has a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical maximum computing capacity of 10^{20} integer or floating-point operations per second for training AI.

(c) Because I find that additional steps must be taken to deal with the national emergency related to significant malicious cyber-enabled activities declared in Executive Order 13694 of April 1, 2015 (Blocking the Property of Certain Persons Engaging in Significant Malicious Cyber-Enabled Activities), as amended by Executive Order 13757 of December 28, 2016 (Taking Additional Steps to Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities), and further amended by Executive Order 13984, to address the use of United

States Infrastructure as a Service (IaaS) Products by foreign malicious cyber actors, including to impose additional record-keeping obligations with respect to foreign transactions and to assist in the investigation of transactions involving foreign malicious cyber actors, I hereby direct the Secretary of Commerce, within 90 days of the date of this order, to:

(i) Propose regulations that require United States IaaS Providers to submit a report to the Secretary of Commerce when a foreign person transacts with that United States IaaS Provider to train a large AI model with potential capabilities that could be used in malicious cyber-enabled activity (a "training run"). Such reports shall include, at a minimum, the identity of the foreign person and the existence of any training run of an AI model meeting the criteria set forth in this section, or other criteria defined by the Secretary in regulations, as well as any additional information identified by the Secretary.

(ii) Include a requirement in the regulations proposed pursuant to subsection 4.2(c)(i) of this section that United States IaaS Providers prohibit any foreign reseller of their United States IaaS Product from providing those products unless such foreign reseller submits to the United States IaaS Provider a report, which the United States IaaS Provider must provide to the Secretary of Commerce, detailing each instance in which a foreign person transacts with the foreign reseller to use the United States IaaS Product to conduct a training run described in subsection 4.2(c)(i) of this section. Such reports shall include, at a minimum, the information specified in subsection 4.2(c)(i) of this section as well as any additional information identified by the Secretary.

(iii) Determine the set of technical conditions for a large AI model to have potential capabilities that could be used in malicious cyber-enabled activity, and revise that determination as necessary and appropriate. Until the Secretary makes such a determination, a model shall

be considered to have potential capabilities that could be used in malicious cyber-enabled activity if it requires a quantity of computing power greater than 10^{26} integer or floating-point operations and is trained on a computing cluster that has a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical maximum compute capacity of 10^{20} integer or floating-point operations per second for training AI.

(d) Within 180 days of the date of this order, pursuant to the finding set forth in subsection 4.2(c) of this section, the Secretary of Commerce shall propose regulations that require United States IaaS Providers to ensure that foreign resellers of United States IaaS Products verify the identity of any foreign person that obtains an IaaS account (account) from the foreign reseller. These regulations shall, at a minimum:

(i) Set forth the minimum standards that a United States IaaS Provider must require of foreign resellers of its United States IaaS Products to verify the identity of a foreign person who opens an account or maintains an existing account with a foreign reseller, including:

(A) the types of documentation and procedures that foreign resellers of United States IaaS Products must require to verify the identity of any foreign person acting as a lessee or sub-lessee of these products or services;

(B) records that foreign resellers of United States IaaS Products must securely maintain regarding a foreign person that obtains an account, including information establishing:

(1) the identity of such foreign person, including name and address;

(2) the means and source of payment (including any associated financial institution and other identifiers such as credit card number, account number, customer identifier, transaction

identifiers, or virtual currency wallet or wallet address identifier);

(3) the electronic mail address and telephonic contact information used to verify a foreign person's identity; and

(4) the Internet Protocol addresses used for access or administration and the date and time of each such access or administrative action related to ongoing verification of such foreign person's ownership of such an account; and

(C) methods that foreign resellers of United States IaaS Products must implement to limit all third-party access to the information described in this subsection, except insofar as such access is otherwise consistent with this order and allowed under applicable law;

(ii) Take into consideration the types of accounts maintained by foreign resellers of United States IaaS Products, methods of opening an account, and types of identifying information available to accomplish the objectives of identifying foreign malicious cyber actors using any such products and avoiding the imposition of an undue burden on such resellers; and

(iii) Provide that the Secretary of Commerce, in accordance with such standards and procedures as the Secretary may delineate and in consultation with the Secretary of Defense, the Attorney General, the Secretary of Homeland Security, and the Director of National Intelligence, may exempt a United States IaaS Provider with respect to any specific foreign reseller of their United States IaaS Products, or with respect to any specific type of account or lessee, from the requirements of any regulation issued pursuant to this subsection. Such standards and procedures may include a finding by the Secretary that such foreign reseller, account, or lessee complies with security best practices to otherwise deter abuse of United States IaaS Products.

(e) The Secretary of Commerce is hereby authorized to take such actions, including the promulgation of rules and regulations, and to employ all powers granted to the President by the International Emergency Economic Powers Act, 50 U.S.C. 1701 *et seq.*, as may be necessary to carry out the purposes of subsections 4.2(c) and (d) of this section. Such actions may include a requirement that United States IaaS Providers require foreign resellers of United States IaaS Products to provide United States IaaS Providers verifications relative to those subsections.

4.3. Managing AI in critical infrastructure and in cybersecurity.

(a) To ensure the protection of critical infrastructure, the following actions shall be taken:

(i) Within 90 days of the date of this order, and at least annually thereafter, the head of each agency with relevant regulatory authority over critical infrastructure and the heads of relevant SRMAs, in coordination with the Director of the Cybersecurity and Infrastructure Security Agency within the Department of Homeland Security for consideration of cross-sector risks, shall evaluate and provide to the Secretary of Homeland Security an assessment of potential risks related to the use of AI in critical infrastructure sectors involved, including ways in which deploying AI may make critical infrastructure systems more vulnerable to critical failures, physical attacks, and cyberattacks, and shall consider ways to mitigate these vulnerabilities. Independent regulatory agencies are encouraged, as they deem appropriate, to contribute to sector-specific risk assessments.

(ii) Within 150 days of the date of this order, the Secretary of the Treasury shall issue a public report on best practices for financial institutions to manage AI-specific cybersecurity risks.

(iii) Within 180 days of the date of this order, the Secretary of Homeland Security, in coordination with the Secretary of Commerce and with SRMAs and other regulators as determined by the Secretary of

Homeland Security, shall incorporate as appropriate the AI Risk Management Framework, NIST AI 100-1, as well as other appropriate security guidance, into relevant safety and security guidelines for use by critical infrastructure owners and operators.

(iv) Within 240 days of the completion of the guidelines described in subsection 4.3(a)(iii) of this section, the Assistant to the President for National Security Affairs and the Director of OMB, in consultation with the Secretary of Homeland Security, shall coordinate work by the heads of agencies with authority over critical infrastructure to develop and take steps for the Federal Government to mandate such guidelines, or appropriate portions thereof, through regulatory or other appropriate action. Independent regulatory agencies are encouraged, as they deem appropriate, to consider whether to mandate guidance through regulatory action in their areas of authority and responsibility.

(v) The Secretary of Homeland Security shall establish an Artificial Intelligence Safety and Security Board as an advisory committee pursuant to section 871 of the Homeland Security Act of 2002 (Public Law 107-296). The Advisory Committee shall include AI experts from the private sector, academia, and government, as appropriate, and provide to the Secretary of Homeland Security and the Federal Government's critical infrastructure community advice, information, or recommendations for improving security, resilience, and incident response related to AI usage in critical infrastructure.

(b) To capitalize on AI's potential to improve United States cyber defenses:

(i) The Secretary of Defense shall carry out the actions described in subsections 4.3(b)(ii) and (iii) of this section for national security systems, and the Secretary of Homeland Security shall carry out these actions for non-national security systems. Each shall do so in

consultation with the heads of other relevant agencies as the Secretary of Defense and the Secretary of Homeland Security may deem appropriate.

(ii) As set forth in subsection 4.3(b)(i) of this section, within 180 days of the date of this order, the Secretary of Defense and the Secretary of Homeland Security shall, consistent with applicable law, each develop plans for, conduct, and complete an operational pilot project to identify, develop, test, evaluate, and deploy AI capabilities, such as large-language models, to aid in the discovery and remediation of vulnerabilities in critical United States Government software, systems, and networks.

(iii) As set forth in subsection 4.3(b)(i) of this section, within 270 days of the date of this order, the Secretary of Defense and the Secretary of Homeland Security shall each provide a report to the Assistant to the President for National Security Affairs on the results of actions taken pursuant to the plans and operational pilot projects required by subsection 4.3(b)(ii) of this section, including a description of any vulnerabilities found and fixed through the development and deployment of AI capabilities and any lessons learned on how to identify, develop, test, evaluate, and deploy AI capabilities effectively for cyber defense.

4.4. Reducing risks at the intersection of AI and CBRN threats.

(a) To better understand and mitigate the risk of AI being misused to assist in the development or use of CBRN threats — with a particular focus on biological weapons — the following actions shall be taken:

(i) Within 180 days of the date of this order, the Secretary of Homeland Security, in consultation with the Secretary of Energy and the Director of the Office of Science and Technology Policy (OSTP), shall evaluate the potential for AI to be misused to enable the development or production of CBRN threats, while also considering the

benefits and application of AI to counter these threats, including, as appropriate, the results of work conducted under section 8(b) of this order. The Secretary of Homeland Security shall:

(A) consult with experts in AI and CBRN issues from the Department of Energy, private AI laboratories, academia, and third-party model evaluators, as appropriate, to evaluate AI model capabilities to present CBRN threats — for the sole purpose of guarding against those threats — as well as options for minimizing the risks of AI model misuse to generate or exacerbate those threats; and

(B) submit a report to the President that describes the progress of these efforts, including an assessment of the types of AI models that may present CBRN risks to the United States, and that makes recommendations for regulating or overseeing the training, deployment, publication, or use of these models, including requirements for safety evaluations and guardrails for mitigating potential threats to national security.

(ii) Within 120 days of the date of this order, the Secretary of Defense, in consultation with the Assistant to the President for National Security Affairs and the Director of OSTP, shall enter into a contract with the National Academies of Sciences, Engineering, and Medicine to conduct — and submit to the Secretary of Defense, the Assistant to the President for National Security Affairs, the Director of the Office of Pandemic Preparedness and Response Policy, the Director of OSTP, and the Chair of the Chief Data Officer Council — a study that:

(A) assesses the ways in which AI can increase biosecurity risks, including risks from generative AI models trained on biological data, and makes recommendations on how to mitigate these risks;

(B) considers the national security implications of the use of data and datasets, especially those associated with pathogens and

omics studies, that the United States Government hosts, generates, funds the creation of, or otherwise owns, for the training of generative AI models, and makes recommendations on how to mitigate the risks related to the use of these data and datasets;

(C) assesses the ways in which AI applied to biology can be used to reduce biosecurity risks, including recommendations on opportunities to coordinate data and high-performance computing resources; and

(D) considers additional concerns and opportunities at the intersection of AI and synthetic biology that the Secretary of Defense deems appropriate.

(b) To reduce the risk of misuse of synthetic nucleic acids, which could be substantially increased by AI's capabilities in this area, and improve biosecurity measures for the nucleic acid synthesis industry, the following actions shall be taken:

(i) Within 180 days of the date of this order, the Director of OSTP, in consultation with the Secretary of State, the Secretary of Defense, the Attorney General, the Secretary of Commerce, the Secretary of Health and Human Services (HHS), the Secretary of Energy, the Secretary of Homeland Security, the Director of National Intelligence, and the heads of other relevant agencies as the Director of OSTP may deem appropriate, shall establish a framework, incorporating, as appropriate, existing United States Government guidance, to encourage providers of synthetic nucleic acid sequences to implement comprehensive, scalable, and verifiable synthetic nucleic acid procurement screening mechanisms, including standards and recommended incentives. As part of this framework, the Director of OSTP shall:

(A) establish criteria and mechanisms for ongoing identification of biological sequences that could be used in a manner that would pose a risk to the national security of the United States; and

(B) determine standardized methodologies and tools for conducting and verifying the performance of sequence synthesis procurement screening, including customer screening approaches to support due diligence with respect to managing security risks posed by purchasers of biological sequences identified in subsection 4.4(b)(i)(A) of this section, and processes for the reporting of concerning activity to enforcement entities.

(ii) Within 180 days of the date of this order, the Secretary of Commerce, acting through the Director of NIST, in coordination with the Director of OSTP, and in consultation with the Secretary of State, the Secretary of HHS, and the heads of other relevant agencies as the Secretary of Commerce may deem appropriate, shall initiate an effort to engage with industry and relevant stakeholders, informed by the framework developed under subsection 4.4(b)(i) of this section, to develop and refine for possible use by synthetic nucleic acid sequence providers:

(A) specifications for effective nucleic acid synthesis procurement screening;

(B) best practices, including security and access controls, for managing sequence-of-concern databases to support such screening;

(C) technical implementation guides for effective screening; and

(D) conformity-assessment best practices and mechanisms.

(iii) Within 180 days of the establishment of the framework pursuant to subsection 4.4(b)(i) of this section, all agencies that fund life-sciences research shall, as appropriate and consistent with applicable law, establish that, as a requirement of funding, synthetic nucleic acid procurement is conducted through providers or manufacturers that adhere to the framework, such as through an attestation from the provider or manufacturer. The Assistant to the President for National

Security Affairs and the Director of OSTP shall coordinate the process of reviewing such funding requirements to facilitate consistency in implementation of the framework across funding agencies.

(iv) In order to facilitate effective implementation of the measures described in subsections 4.4(b)(i)-(iii) of this section, the Secretary of Homeland Security, in consultation with the heads of other relevant agencies as the Secretary of Homeland Security may deem appropriate, shall:

(A) within 180 days of the establishment of the framework pursuant to subsection 4.4(b)(i) of this section, develop a framework to conduct structured evaluation and stress testing of nucleic acid synthesis procurement screening, including the systems developed in accordance with subsections 4.4(b)(i)-(ii) of this section and implemented by providers of synthetic nucleic acid sequences; and

(B) following development of the framework pursuant to subsection 4.4(b)(iv)(A) of this section, submit an annual report to the Assistant to the President for National Security Affairs, the Director of the Office of Pandemic Preparedness and Response Policy, and the Director of OSTP on any results of the activities conducted pursuant to subsection 4.4(b)(iv)(A) of this section, including recommendations, if any, on how to strengthen nucleic acid synthesis procurement screening, including customer screening systems.

4.5. Reducing the risks posed by synthetic content

To foster capabilities for identifying and labelling synthetic content produced by AI systems, and to establish the authenticity and provenance of digital content, both synthetic and not synthetic, produced by the Federal Government or on its behalf:

(a) Within 240 days of the date of this order, the Secretary of Commerce, in consultation with the heads of other relevant agencies as the Secretary of Commerce may deem appropriate, shall submit a report to the Director of OMB and the Assistant to the President for National Security Affairs identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, for:

(i) authenticating content and tracking its provenance;

(ii) labeling synthetic content, such as using watermarking;

(iii) detecting synthetic content;

(iv) preventing generative AI from producing child sexual abuse material or producing non-consensual intimate imagery of real individuals (to include intimate digital depictions of the body or body parts of an identifiable individual);

(v) testing software used for the above purposes; and

(vi) auditing and maintaining synthetic content.

(b) Within 180 days of submitting the report required under subsection 4.5(a) of this section, and updated periodically thereafter, the Secretary of Commerce, in coordination with the Director of OMB, shall develop guidance regarding the existing tools and practices for digital content authentication and synthetic content detection measures. The guidance shall include measures for the purposes listed in subsection 4.5(a) of this section.

(c) Within 180 days of the development of the guidance required under subsection 4.5(b) of this section, and updated periodically thereafter, the Director of OMB, in consultation with the Secretary of State; the Secretary of Defense; the Attorney General; the Secretary of Commerce, acting through the Director of NIST; the Secretary of Homeland Security; the Director of National Intelligence; and the heads of other agencies that the Director of OMB deems appropriate, shall — for the purpose of

strengthening public confidence in the integrity of official United States Government digital content — issue guidance to agencies for labeling and authenticating such content that they produce or publish.

(d) The Federal Acquisition Regulatory Council shall, as appropriate and consistent with applicable law, consider amending the Federal Acquisition Regulation to take into account the guidance established under subsection 4.5 of this section.

4.6. Soliciting input on dual-use foundation models with widely available model weights

When the weights for a dual-use foundation model are widely available — such as when they are publicly posted on the Internet — there can be substantial benefits to innovation, but also substantial security risks, such as the removal of safeguards within the model. To address the risks and potential benefits of dual-use foundation models with widely available weights, within 270 days of the date of this order, the Secretary of Commerce, acting through the Assistant Secretary of Commerce for Communications and Information, and in consultation with the Secretary of State, shall:

(a) solicit input from the private sector, academia, civil society, and other stakeholders through a public consultation process on potential risks, benefits, other implications, and appropriate policy and regulatory approaches related to dual-use foundation models for which the model weights are widely available, including:

(i) risks associated with actors fine-tuning dual-use foundation models for which the model weights are widely available or removing those models' safeguards;

(ii) benefits to AI innovation and research, including research into AI safety and risk management, of dual-use foundation models for which the model weights are widely available; and

(iii) potential voluntary, regulatory, and international mechanisms to manage the risks and maximize the benefits of dual-use foundation models for which the model weights are widely available; and

(b) based on input from the process described in subsection 4.6(a) of this section, and in consultation with the heads of other relevant agencies as the Secretary of Commerce deems appropriate, submit a report to the President on the potential benefits, risks, and implications of dual-use foundation models for which the model weights are widely available, as well as policy and regulatory recommendations pertaining to those models.

4.7. Promoting safe release and preventing the malicious use of federal data for AI training

To improve public data access and manage security risks, and consistent with the objectives of the Open, Public, Electronic, and Necessary Government Data Act (title II of Public Law 115-435) to expand public access to Federal data assets in a machine-readable format while also taking into account security considerations, including the risk that information in an individual data asset in isolation does not pose a security risk but, when combined with other available information, may pose such a risk:

(a) within 270 days of the date of this order, the Chief Data Officer Council, in consultation with the Secretary of Defense, the Secretary of Commerce, the Secretary of Energy, the Secretary of Homeland Security, and the Director of National Intelligence, shall develop initial guidelines for performing security reviews, including reviews to identify and manage the potential security risks of releasing Federal data that could aid in the development of CBRN weapons as well as the development of autonomous offensive cyber capabilities, while also providing public access to Federal Government data in line with the goals stated in the Open, Public,

Electronic, and Necessary Government Data Act (title II of Public Law 115-435); and

(b) within 180 days of the development of the initial guidelines required by subsection 4.7(a) of this section, agencies shall conduct a security review of all data assets in the comprehensive data inventory required under 44 U.S.C. 3511(a)(1) and (2)(B) and shall take steps, as appropriate and consistent with applicable law, to address the highest-priority potential security risks that releasing that data could raise with respect to CBRN weapons, such as the ways in which that data could be used to train AI systems.

4.8. Directing the development of a national security memorandum

To develop a coordinated executive branch approach to managing AI's security risks, the Assistant to the President for National Security Affairs and the Assistant to the President and Deputy Chief of Staff for Policy shall oversee an interagency process with the purpose of, within 270 days of the date of this order, developing and submitting a proposed National Security Memorandum on AI to the President. The memorandum shall address the governance of AI used as a component of a national security system or for military and intelligence purposes. The memorandum shall take into account current efforts to govern the development and use of AI for national security systems. The memorandum shall outline actions for the Department of Defense, the Department of State, other relevant agencies, and the Intelligence Community to address the national security risks and potential benefits posed by AI. In particular, the memorandum shall:

(a) provide guidance to the Department of Defense, other relevant agencies, and the Intelligence Community on the continued adoption of AI capabilities to advance the United States national security mission, including through directing specific AI assurance and risk-management practices for national security uses of AI that may affect the rights or

safety of United States persons and, in appropriate contexts, non-United States persons; and

(b) direct continued actions, as appropriate and consistent with applicable law, to address the potential use of AI systems by adversaries and other foreign actors in ways that threaten the capabilities or objectives of the Department of Defense or the Intelligence Community, or that otherwise pose risks to the security of the United States or its allies and partners.

22.2.5 Section 5. Promoting innovation and competition

5.1 Attracting AI talent to the United States

(a) Within 90 days of the date of this order, to attract and retain talent in AI and other critical and emerging technologies in the United States economy, the Secretary of State and the Secretary of Homeland Security shall take appropriate steps to:

(i) streamline processing times of visa petitions and applications, including by ensuring timely availability of visa appointments, for noncitizens who seek to travel to the United States to work on, study, or conduct research in AI or other critical and emerging technologies; and

(ii) facilitate continued availability of visa appointments in sufficient volume for applicants with expertise in AI or other critical and emerging technologies.

(b) Within 120 days of the date of this order, the Secretary of State shall:

(i) consider initiating a rulemaking to establish new criteria to designate countries and skills on the Department of State's Exchange Visitor Skills List as it relates to the 2-year foreign residence requirement for certain J-1 non-immigrants, including those skills that are critical to the United States;

(ii) consider publishing updates to the 2009 Revised Exchange Visitor Skills List (74 FR 20108); and

(iii) consider implementing a domestic visa renewal program under 22 C.F.R. 41.111(b) to facilitate the ability of qualified applicants, including highly skilled talent in AI and critical and emerging technologies, to continue their work in the United States without unnecessary interruption.

(c) Within 180 days of the date of this order, the Secretary of State shall:

(i) consider initiating a rulemaking to expand the categories of non-immigrants who qualify for the domestic visa renewal program covered under 22 C.F.R. 41.111(b) to include academic J-1 research scholars and F-1 students in science, technology, engineering, and mathematics (STEM); and

(ii) establish, to the extent permitted by law and available appropriations, a program to identify and attract top talent in AI and other critical and emerging technologies at universities, research institutions, and the private sector overseas, and to establish and increase connections with that talent to educate them on opportunities and resources for research and employment in the United States, including overseas educational components to inform top STEM talent of non-immigrant and immigrant visa options and potential expedited adjudication of their visa petitions and applications.

(d) Within 180 days of the date of this order, the Secretary of Homeland Security shall:

(i) review and initiate any policy changes the Secretary determines necessary and appropriate to clarify and modernize immigration pathways for experts in AI and other critical and emerging technologies, including O-1A and EB-1 noncitizens of extraordinary ability; EB-2 advanced-degree holders and noncitizens of exceptional ability; and

startup founders in AI and other critical and emerging technologies using the International Entrepreneur Rule; and

(ii) continue its rulemaking process to modernize the H-1B program and enhance its integrity and usage, including by experts in AI and other critical and emerging technologies, and consider initiating a rulemaking to enhance the process for noncitizens, including experts in AI and other critical and emerging technologies and their spouses, dependents, and children, to adjust their status to lawful permanent resident.

(e) Within 45 days of the date of this order, for purposes of considering updates to the "Schedule A" list of occupations, 20 C.F.R. 656.5, the Secretary of Labor shall publish a request for information (RFI) to solicit public input, including from industry and worker-advocate communities, identifying AI and other STEM-related occupations, as well as additional occupations across the economy, for which there is an insufficient number of ready, willing, able, and qualified United States workers.

(f) The Secretary of State and the Secretary of Homeland Security shall, consistent with applicable law and implementing regulations, use their discretionary authorities to support and attract foreign nationals with special skills in AI and other critical and emerging technologies seeking to work, study, or conduct research in the United States.

(g) Within 120 days of the date of this order, the Secretary of Homeland Security, in consultation with the Secretary of State, the Secretary of Commerce, and the Director of OSTP, shall develop and publish informational resources to better attract and retain experts in AI and other critical and emerging technologies, including:

(i) a clear and comprehensive guide for experts in AI and other critical and emerging technologies to understand their options for working in the United States, to be published in multiple relevant languages on AI.gov; and

(ii) a public report with relevant data on applications, petitions, approvals, and other key indicators of how experts in AI and other critical and emerging technologies have utilized the immigration system through the end of Fiscal Year 2023.

5.2 Promoting innovation

(a) To develop and strengthen public-private partnerships for advancing innovation, commercialization, and risk-mitigation methods for AI, and to help promote safe, responsible, fair, privacy-protecting, and trustworthy AI systems, the Director of NSF shall take the following steps:

(i) Within 90 days of the date of this order, in coordination with the heads of agencies that the Director of NSF deems appropriate, launch a pilot program implementing the National AI Research Resource (NAIRR), consistent with past recommendations of the NAIRR Task Force. The program shall pursue the infrastructure, governance mechanisms, and user interfaces to pilot an initial integration of distributed computational, data, model, and training resources to be made available to the research community in support of AI-related research and development. The Director of NSF shall identify Federal and private sector computational, data, software, and training resources appropriate for inclusion in the NAIRR pilot program. To assist with such work, within 45 days of the date of this order, the heads of agencies whom the Director of NSF identifies for coordination pursuant to this subsection shall each submit to the Director of NSF a report identifying the agency resources that could be developed and integrated into such a pilot program. These reports shall include a description of such resources, including their current status and availability; their format, structure, or technical specifications; associated agency expertise that will be provided; and the benefits and risks associated with their inclusion in the NAIRR pilot program. The heads of independent

regulatory agencies are encouraged to take similar steps, as they deem appropriate.

(ii) Within 150 days of the date of this order, fund and launch at least one NSF Regional Innovation Engine that prioritizes AI-related work, such as AI-related research, societal, or workforce needs.

(iii) Within 540 days of the date of this order, establish at least four new National AI Research Institutes, in addition to the 25 currently funded as of the date of this order.

(b) Within 120 days of the date of this order, to support activities involving high-performance and data-intensive computing, the Secretary of Energy, in coordination with the Director of NSF, shall, in a manner consistent with applicable law and available appropriations, establish a pilot program to enhance existing successful training programs for scientists, with the goal of training 500 new researchers by 2025 capable of meeting the rising demand for AI talent.

(c) To promote innovation and clarify issues related to AI and inventorship of patentable subject matter, the Under Secretary of Commerce for Intellectual Property and Director of the United States Patent and Trademark Office (USPTO Director) shall:

(i) within 120 days of the date of this order, publish guidance to USPTO patent examiners and applicants addressing inventorship and the use of AI, including generative AI, in the inventive process, including illustrative examples in which AI systems play different roles in inventive processes and how, in each example, inventorship issues ought to be analysed;

(ii) subsequently, within 270 days of the date of this order, issue additional guidance to USPTO patent examiners and applicants to address other considerations at the intersection of AI and IP, which could include, as the USPTO Director deems necessary, updated

guidance on patent eligibility to address innovation in AI and critical and emerging technologies; and

(iii) within 270 days of the date of this order or 180 days after the United States Copyright Office of the Library of Congress publishes its forthcoming AI study that will address copyright issues raised by AI, whichever comes later, consult with the Director of the United States Copyright Office and issue recommendations to the President on potential executive actions relating to copyright and AI. The recommendations shall address any copyright and related issues discussed in the United States Copyright Office's study, including the scope of protection for works produced using AI and the treatment of copyrighted works in AI training.

(d) Within 180 days of the date of this order, to assist developers of AI in combatting AI-related IP risks, the Secretary of Homeland Security, acting through the Director of the National Intellectual Property Rights Coordination Center, and in consultation with the Attorney General, shall develop a training, analysis, and evaluation program to mitigate AI-related IP risks. Such a program shall:

(i) include appropriate personnel dedicated to collecting and analyzing reports of AI-related IP theft, investigating such incidents with implications for national security, and, where appropriate and consistent with applicable law, pursuing related enforcement actions;

(ii) implement a policy of sharing information and coordinating on such work, as appropriate and consistent with applicable law, with the Federal Bureau of Investigation; United States Customs and Border Protection; other agencies; State and local agencies; and appropriate international organizations, including through work-sharing agreements;

(iii) develop guidance and other appropriate resources to assist private sector actors with mitigating the risks of AI-related IP theft;

(iv) share information and best practices with AI developers and law enforcement personnel to identify incidents, inform stakeholders of current legal requirements, and evaluate AI systems for IP law violations, as well as develop mitigation strategies and resources; and

(v) assist the Intellectual Property Enforcement Coordinator in updating the Intellectual Property Enforcement Coordinator Joint Strategic Plan on Intellectual Property Enforcement to address AI-related issues.

(e) To advance responsible AI innovation by a wide range of healthcare technology developers that promotes the welfare of patients and workers in the healthcare sector, the Secretary of HHS shall identify and, as appropriate and consistent with applicable law and the activities directed in section 8 of this order, prioritize grantmaking and other awards, as well as undertake related efforts, to support responsible AI development and use, including:

(i) collaborating with appropriate private sector actors through HHS programs that may support the advancement of AI-enabled tools that develop personalized immune-response profiles for patients, consistent with section 4 of this order;

(ii) prioritizing the allocation of 2024 Leading Edge Acceleration Project cooperative agreement awards to initiatives that explore ways to improve healthcare-data quality to support the responsible development of AI tools for clinical care, real-world-evidence programs, population health, public health, and related research; and

(iii) accelerating grants awarded through the National Institutes of Health Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program and showcasing current AIM-AHEAD activities in underserved communities.

(f) To advance the development of AI systems that improve the quality of veterans' healthcare, and in order to support small businesses' innovative capacity, the Secretary of Veterans Affairs shall:

(i) within 365 days of the date of this order, host two 3-month nationwide AI Tech Sprint competitions; and

(ii) as part of the AI Tech Sprint competitions and in collaboration with appropriate partners, provide participants access to technical assistance, mentorship opportunities, individualized expert feedback on products under development, potential contract opportunities, and other programming and resources.

(g) Within 180 days of the date of this order, to support the goal of strengthening our Nation's resilience against climate change impacts and building an equitable clean energy economy for the future, the Secretary of Energy, in consultation with the Chair of the Federal Energy Regulatory Commission, the Director of OSTP, the Chair of the Council on Environmental Quality, the Assistant to the President and National Climate Advisor, and the heads of other relevant agencies as the Secretary of Energy may deem appropriate, shall:

(i) issue a public report describing the potential for AI to improve planning, permitting, investment, and operations for electric grid infrastructure and to enable the provision of clean, affordable, reliable, resilient, and secure electric power to all Americans;

(ii) develop tools that facilitate building foundation models useful for basic and applied science, including models that streamline permitting and environmental reviews while improving environmental and social outcomes;

(iii) collaborate, as appropriate, with private sector organizations and members of academia to support development of AI tools to mitigate climate change risks;

(iv) take steps to expand partnerships with industry, academia, other agencies, and international allies and partners to utilize the Department of Energy's computing capabilities and AI testbeds to build foundation models that support new applications in science and energy, and for national security, including partnerships that increase community preparedness for climate-related risks, enable clean-energy deployment (including addressing delays in permitting reviews), and enhance grid reliability and resilience; and

(v) establish an office to coordinate development of AI and other critical and emerging technologies across Department of Energy programs and the 17 National Laboratories.

(h) Within 180 days of the date of this order, to understand AI's implications for scientific research, the President's Council of Advisors on Science and Technology shall submit to the President and make publicly available a report on the potential role of AI, especially given recent developments in AI, in research aimed at tackling major societal and global challenges. The report shall include a discussion of issues that may hinder the effective use of AI in research and practices needed to ensure that AI is used responsibly for research.

5.3 Promoting competition

(a) The head of each agency developing policies and regulations related to AI shall use their authorities, as appropriate and consistent with applicable law, to promote competition in AI and related technologies, as well as in other markets. Such actions include addressing risks arising from concentrated control of key inputs, taking steps to stop unlawful collusion and prevent dominant firms from disadvantaging competitors, and working to provide new opportunities for small businesses and entrepreneurs. In particular, the Federal Trade Commission is encouraged to consider, as it deems appropriate, whether to exercise the Commission's existing authorities, including its rulemaking authority under the

Federal Trade Commission Act, 15 U.S.C. 41 et seq., to ensure fair competition in the AI marketplace and to ensure that consumers and workers are protected from harms that may be enabled by the use of AI.

(b) To promote competition and innovation in the semiconductor industry, recognizing that semiconductors power AI technologies and that their availability is critical to AI competition, the Secretary of Commerce shall, in implementing division A of Public Law 117-167, known as the Creating Helpful Incentives to Produce Semiconductors (CHIPS) Act of 2022, promote competition by:

(i) implementing a flexible membership structure for the National Semiconductor Technology Center that attracts all parts of the semiconductor and microelectronics ecosystem, including startups and small firms;

(ii) implementing mentorship programs to increase interest and participation in the semiconductor industry, including from workers in underserved communities;

(iii) increasing, where appropriate and to the extent permitted by law, the availability of resources to startups and small businesses, including:

(A) funding for physical assets, such as specialty equipment or facilities, to which startups and small businesses may not otherwise have access;

(B) datasets — potentially including test and performance data — collected, aggregated, or shared by CHIPS research and development programs;

(C) workforce development programs;

(D) design and process technology, as well as IP, as appropriate; and

(E) other resources, including technical and intellectual property assistance, that could accelerate commercialization of new technologies by startups and small businesses, as appropriate; and

(iv) considering the inclusion, to the maximum extent possible, and as consistent with applicable law, of competition-increasing measures in notices of funding availability for commercial research-and-development facilities focused on semiconductors, including measures that increase access to facility capacity for startups or small firms developing semiconductors used to power AI technologies.

(c) To support small businesses innovating and commercializing AI, as well as in responsibly adopting and deploying AI, the Administrator of the Small Business Administration shall:

(i) prioritize the allocation of Regional Innovation Cluster program funding for clusters that support planning activities related to the establishment of one or more Small Business AI Innovation and Commercialization Institutes that provide support, technical assistance, and other resources to small businesses seeking to innovate, commercialize, scale, or otherwise advance the development of AI;

(ii) prioritize the allocation of up to $2 million in Growth Accelerator Fund Competition bonus prize funds for accelerators that support the incorporation or expansion of AI-related curricula, training, and technical assistance, or other AI-related resources within their programming; and

(iii) assess the extent to which the eligibility criteria of existing programs, including the State Trade Expansion Program, Technical and Business Assistance funding, and capital-access programs — such as the 7(a) loan program, 504 loan program, and Small Business Investment Company (SBIC) program — support appropriate expenses by small businesses related to the adoption of AI and, if feasible and

appropriate, revise eligibility criteria to improve support for these expenses.

(d) The Administrator of the Small Business Administration, in coordination with resource partners, shall conduct outreach regarding, and raise awareness of, opportunities for small businesses to use capital-access programs described in subsection 5.3(c) of this section for eligible AI-related purposes, and for eligible investment funds with AI-related expertise — particularly those seeking to serve or with experience serving underserved communities — to apply for an SBIC license.

22.2.6 Section 6. Supporting workers

(a) To advance the Government's understanding of AI's implications for workers, the following actions shall be taken within 180 days of the date of this order:

(i) The Chairman of the Council of Economic Advisers shall prepare and submit a report to the President on the labor-market effects of AI.

(ii) To evaluate necessary steps for the Federal Government to address AI-related workforce disruptions, the Secretary of Labor shall submit to the President a report analyzing the abilities of agencies to support workers displaced by the adoption of AI and other technological advancements. The report shall, at a minimum:

(A) assess how current or formerly operational Federal programs designed to assist workers facing job disruptions — including unemployment insurance and programs authorized by the Workforce Innovation and Opportunity Act (Public Law 113-128) — could be used to respond to possible future AI-related disruptions; and

(B) identify options, including potential legislative measures, to strengthen or develop additional Federal support for workers displaced by AI and, in consultation with the Secretary of Commerce and the Secretary of Education, strengthen and expand education

and training opportunities that provide individuals pathways to occupations related to AI.

(b) To help ensure that AI deployed in the workplace advances employees' well-being:

(i) The Secretary of Labor shall, within 180 days of the date of this order and in consultation with other agencies and with outside entities, including labor unions and workers, as the Secretary of Labor deems appropriate, develop and publish principles and best practices for employers that could be used to mitigate AI's potential harms to employees' well-being and maximize its potential benefits. The principles and best practices shall include specific steps for employers to take with regard to AI, and shall cover, at a minimum:

(A) job-displacement risks and career opportunities related to AI, including effects on job skills and evaluation of applicants and workers;

(B) labor standards and job quality, including issues related to the equity, protected-activity, compensation, health, and safety implications of AI in the workplace; and

(C) implications for workers of employers' AI-related collection and use of data about them, including transparency, engagement, management, and activity protected under worker-protection laws.

(ii) After principles and best practices are developed pursuant to subsection (b)(i) of this section, the heads of agencies shall consider, in consultation with the Secretary of Labor, encouraging the adoption of these guidelines in their programs to the extent appropriate for each program and consistent with applicable law.

(iii) To support employees whose work is monitored or augmented by AI in being compensated appropriately for all of their work time, the Secretary of Labor shall issue guidance to make clear that

employers that deploy AI to monitor or augment employees' work must continue to comply with protections that ensure that workers are compensated for their hours worked, as defined under the Fair Labor Standards Act of 1938, 29 U.S.C. 201 *et seq.*, and other legal requirements.

(c) To foster a diverse AI-ready workforce, the Director of NSF shall prioritize available resources to support AI-related education and AI-related workforce development through existing programs. The Director shall additionally consult with agencies, as appropriate, to identify further opportunities for agencies to allocate resources for those purposes. The actions by the Director shall use appropriate fellowship programs and awards for these purposes.

22.2.7 Section 7. Advancing equity and civil rights

7.1. Strengthening AI and civil rights in the criminal justice system

(a) To address unlawful discrimination and other harms that may be exacerbated by AI, the Attorney General shall:

(i) consistent with Executive Order 12250 of November 2, 1980 (Leadership and Coordination of Non-discrimination Laws), Executive Order 14091, and 28 C.F.R. 0.50-51, coordinate with and support agencies in their implementation and enforcement of existing Federal laws to address civil rights and civil liberties violations and discrimination related to AI;

(ii) direct the Assistant Attorney General in charge of the Civil Rights Division to convene, within 90 days of the date of this order, a meeting of the heads of Federal civil rights offices — for which meeting the heads of civil rights offices within independent regulatory agencies will be encouraged to join — to discuss comprehensive use of their respective authorities and offices to: prevent and address discrimination in the use of automated systems, including algorithmic

discrimination; increase coordination between the Department of Justice's Civil Rights Division and Federal civil rights offices concerning issues related to AI and algorithmic discrimination; improve external stakeholder engagement to promote public awareness of potential discriminatory uses and effects of AI; and develop, as appropriate, additional training, technical assistance, guidance, or other resources; and

(iii) consider providing, as appropriate and consistent with applicable law, guidance, technical assistance, and training to State, local, Tribal, and territorial investigators and prosecutors on best practices for investigating and prosecuting civil rights violations and discrimination related to automated systems, including AI.

(b) To promote the equitable treatment of individuals and adhere to the Federal Government's fundamental obligation to ensure fair and impartial justice for all, with respect to the use of AI in the criminal justice system, the Attorney General shall, in consultation with the Secretary of Homeland Security and the Director of OSTP:

(i) within 365 days of the date of this order, submit to the President a report that addresses the use of AI in the criminal justice system, including any use in:

(A) sentencing;

(B) parole, supervised release, and probation;

(C) bail, pretrial release, and pretrial detention;

(D) risk assessments, including pretrial, earned time, and early release or transfer to home-confinement determinations;

(E) police surveillance;

(F) crime forecasting and predictive policing, including the ingestion of historical crime data into AI systems to predict high-density "hot spots";

(G) prison-management tools; and

(H) forensic analysis;

(ii) within the report set forth in subsection 7.1(b)(i) of this section:

(A) identify areas where AI can enhance law enforcement efficiency and accuracy, consistent with protections for privacy, civil rights, and civil liberties; and

(B) recommend best practices for law enforcement agencies, including safeguards and appropriate use limits for AI, to address the concerns set forth in section 13(e)(i) of Executive Order 14074 as well as the best practices and the guidelines set forth in section 13(e)(iii) of Executive Order 14074; and

(iii) supplement the report set forth in subsection 7.1(b)(i) of this section as appropriate with recommendations to the President, including with respect to requests for necessary legislation.

(c) To advance the presence of relevant technical experts and expertise (such as machine-learning engineers, software and infrastructure engineering, data privacy experts, data scientists, and user experience researchers) among law enforcement professionals:

(i) The interagency working group created pursuant to section 3 of Executive Order 14074 shall, within 180 days of the date of this order, identify and share best practices for recruiting and hiring law enforcement professionals who have the technical skills mentioned in subsection 7.1(c) of this section, and for training law enforcement professionals about responsible application of AI.

(ii) Within 270 days of the date of this order, the Attorney General shall, in consultation with the Secretary of Homeland Security, consider those best practices and the guidance developed under section 3(d) of Executive Order 14074 and, if necessary, develop additional general recommendations for State, local, Tribal, and territorial law enforcement agencies and criminal justice agencies seeking to recruit, hire, train, promote, and retain highly qualified and service-oriented

officers and staff with relevant technical knowledge. In considering this guidance, the Attorney General shall consult with State, local, Tribal, and territorial law enforcement agencies, as appropriate.

(iii) Within 365 days of the date of this order, the Attorney General shall review the work conducted pursuant to section 2(b) of Executive Order 14074 and, if appropriate, reassess the existing capacity to investigate law enforcement deprivation of rights under color of law resulting from the use of AI, including through improving and increasing training of Federal law enforcement officers, their supervisors, and Federal prosecutors on how to investigate and prosecute cases related to AI involving the deprivation of rights under color of law pursuant to 18 U.S.C. 242.

7.2. Protecting civil rights related to government benefits and programs

(a) To advance equity and civil rights, consistent with the directives of Executive Order 14091, and in addition to complying with the guidance on Federal Government use of AI issued pursuant to section 10.1(b) of this order, agencies shall use their respective civil rights and civil liberties offices and authorities — as appropriate and consistent with applicable law — to prevent and address unlawful discrimination and other harms that result from uses of AI in Federal Government programs and benefits administration. This directive does not apply to agencies' civil or criminal enforcement authorities. Agencies shall consider opportunities to ensure that their respective civil rights and civil liberties offices are appropriately consulted on agency decisions regarding the design, development, acquisition, and use of AI in Federal Government programs and benefits administration. To further these objectives, agencies shall also consider opportunities to increase coordination, communication, and engagement about AI as appropriate with community-based organizations; civil-rights and civil-liberties organizations; academic institutions;

industry; State, local, Tribal, and territorial governments; and other stakeholders.

(b) To promote equitable administration of public benefits:

(i) The Secretary of HHS shall, within 180 days of the date of this order and in consultation with relevant agencies, publish a plan, informed by the guidance issued pursuant to section 10.1(b) of this order, addressing the use of automated or algorithmic systems in the implementation by States and localities of public benefits and services administered by the Secretary, such as to promote: assessment of access to benefits by qualified recipients; notice to recipients about the presence of such systems; regular evaluation to detect unjust denials; processes to retain appropriate levels of discretion of expert agency staff; processes to appeal denials to human reviewers; and analysis of whether algorithmic systems in use by benefit programs achieve equitable and just outcomes.

(ii) The Secretary of Agriculture shall, within 180 days of the date of this order and as informed by the guidance issued pursuant to section 10.1(b) of this order, issue guidance to State, local, Tribal, and territorial public-benefits administrators on the use of automated or algorithmic systems in implementing benefits or in providing customer support for benefit programs administered by the Secretary, to ensure that programs using those systems:

(A) maximize program access for eligible recipients;

(B) employ automated or algorithmic systems in a manner consistent with any requirements for using merit systems personnel in public-benefits programs;

(C) identify instances in which reliance on automated or algorithmic systems would require notification by the State, local, Tribal, or territorial government to the Secretary;

(D) identify instances when applicants and participants can appeal benefit determinations to a human reviewer for reconsideration and can receive other customer support from a human being;

(E) enable auditing and, if necessary, remediation of the logic used to arrive at an individual decision or determination to facilitate the evaluation of appeals; and

(F) enable the analysis of whether algorithmic systems in use by benefit programs achieve equitable outcomes.

7.3. Strengthening AI and civil rights in the broader economy.

(a) Within 365 days of the date of this order, to prevent unlawful discrimination from AI used for hiring, the Secretary of Labor shall publish guidance for Federal contractors regarding non-discrimination in hiring involving AI and other technology-based hiring systems.

(b) To address discrimination and biases against protected groups in housing markets and consumer financial markets, the Director of the Federal Housing Finance Agency and the Director of the Consumer Financial Protection Bureau are encouraged to consider using their authorities, as they deem appropriate, to require their respective regulated entities, where possible, to use appropriate methodologies including AI tools to ensure compliance with Federal law and:

(i) evaluate their underwriting models for bias or disparities affecting protected groups; and

(ii) evaluate automated collateral-valuation and appraisal processes in ways that minimize bias.

(c) Within 180 days of the date of this order, to combat unlawful discrimination enabled by automated or algorithmic tools used to make decisions about access to housing and in other real estate-related transactions, the Secretary of Housing and Urban Development shall, and the Director of the Consumer Financial Protection Bureau is encouraged to, issue additional guidance:

(i) addressing the use of tenant screening systems in ways that may violate the Fair Housing Act (Public Law 90-284), the Fair Credit Reporting Act (Public Law 91-508), or other relevant Federal laws, including how the use of data, such as criminal records, eviction records, and credit information, can lead to discriminatory outcomes in violation of Federal law; and

(ii) addressing how the Fair Housing Act, the Consumer Financial Protection Act of 2010 (title X of Public Law 111-203), or the Equal Credit Opportunity Act (Public Law 93-495) apply to the advertising of housing, credit, and other real estate-related transactions through digital platforms, including those that use algorithms to facilitate advertising delivery, as well as on best practices to avoid violations of Federal law.

(d) To help ensure that people with disabilities benefit from AI's promise while being protected from its risks, including unequal treatment from the use of biometric data like gaze direction, eye tracking, gait analysis, and hand motions, the Architectural and Transportation Barriers Compliance Board is encouraged, as it deems appropriate, to solicit public participation and conduct community engagement; to issue technical assistance and recommendations on the risks and benefits of AI in using biometric data as an input; and to provide people with disabilities access to information and communication technology and transportation services.

22.2.8 Section 8. Protecting consumers, patients, passengers, and students

(a) Independent regulatory agencies are encouraged, as they deem appropriate, to consider using their full range of authorities to protect American consumers from fraud, discrimination, and threats to privacy and to address other risks that may arise from the use of AI, including risks to financial stability, and to consider rulemaking, as well as

emphasizing or clarifying where existing regulations and guidance apply to AI, including clarifying the responsibility of regulated entities to conduct due diligence on and monitor any third-party AI services they use, and emphasizing or clarifying requirements and expectations related to the transparency of AI models and regulated entities' ability to explain their use of AI models.

(b) To help ensure the safe, responsible deployment and use of AI in the healthcare, public-health, and human-services sectors:

(i) Within 90 days of the date of this order, the Secretary of HHS shall, in consultation with the Secretary of Defense and the Secretary of Veterans Affairs, establish an HHS AI Task Force that shall, within 365 days of its creation, develop a strategic plan that includes policies and frameworks — possibly including regulatory action, as appropriate — on responsible deployment and use of AI and AI-enabled technologies in the health and human services sector (including research and discovery, drug and device safety, healthcare delivery and financing, and public health), and identify appropriate guidance and resources to promote that deployment, including in the following areas:

(A) development, maintenance, and use of predictive and generative AI-enabled technologies in healthcare delivery and financing — including quality measurement, performance improvement, program integrity, benefits administration, and patient experience — taking into account considerations such as appropriate human oversight of the application of AI-generated output;

(B) long-term safety and real-world performance monitoring of AI-enabled technologies in the health and human services sector, including clinically relevant or significant modifications and performance across population groups, with a means to communicate product updates to regulators, developers, and users;

(C) incorporation of equity principles in AI-enabled technologies used in the health and human services sector, using disaggregated data on affected populations and representative population data sets when developing new models, monitoring algorithmic performance against discrimination and bias in existing models, and helping to identify and mitigate discrimination and bias in current systems;

(D) incorporation of safety, privacy, and security standards into the software-development lifecycle for protection of personally identifiable information, including measures to address AI-enhanced cybersecurity threats in the health and human services sector;

(E) development, maintenance, and availability of documentation to help users determine appropriate and safe uses of AI in local settings in the health and human services sector;

(F) work to be done with State, local, Tribal, and territorial health and human services agencies to advance positive use cases and best practices for use of AI in local settings; and

(G) identification of uses of AI to promote workplace efficiency and satisfaction in the health and human services sector, including reducing administrative burdens.

(ii) Within 180 days of the date of this order, the Secretary of HHS shall direct HHS components, as the Secretary of HHS deems appropriate, to develop a strategy, in consultation with relevant agencies, to determine whether AI-enabled technologies in the health and human services sector maintain appropriate levels of quality, including, as appropriate, in the areas described in subsection (b)(i) of this section. This work shall include the development of AI assurance policy — to evaluate important aspects of the performance of AI-enabled healthcare tools — and infrastructure needs for enabling pre-market

assessment and post-market oversight of AI-enabled healthcare-technology algorithmic system performance against real-world data.

(iii) Within 180 days of the date of this order, the Secretary of HHS shall, in consultation with relevant agencies as the Secretary of HHS deems appropriate, consider appropriate actions to advance the prompt understanding of, and compliance with, Federal non-discrimination laws by health and human services providers that receive Federal financial assistance, as well as how those laws relate to AI. Such actions may include:

(A) convening and providing technical assistance to health and human services providers and payers about their obligations under Federal non-discrimination and privacy laws as they relate to AI and the potential consequences of noncompliance; and

(B) issuing guidance, or taking other action as appropriate, in response to any complaints or other reports of noncompliance with Federal non-discrimination and privacy laws as they relate to AI.

(iv) Within 365 days of the date of this order, the Secretary of HHS shall, in consultation with the Secretary of Defense and the Secretary of Veterans Affairs, establish an AI safety program that, in partnership with voluntary federally listed Patient Safety Organizations:

(A) establishes a common framework for approaches to identifying and capturing clinical errors resulting from AI deployed in healthcare settings as well as specifications for a central tracking repository for associated incidents that cause harm, including through bias or discrimination, to patients, caregivers, or other parties;

(B) analyses captured data and generated evidence to develop, wherever appropriate, recommendations, best practices, or other informal guidelines aimed at avoiding these harms; and

(C) disseminates those recommendations, best practices, or other informal guidance to appropriate stakeholders, including healthcare providers.

(v) Within 365 days of the date of this order, the Secretary of HHS shall develop a strategy for regulating the use of AI or AI-enabled tools in drug-development processes. The strategy shall, at a minimum:

(A) define the objectives, goals, and high-level principles required for appropriate regulation throughout each phase of drug development;

(B) identify areas where future rulemaking, guidance, or additional statutory authority may be necessary to implement such a regulatory system;

(C) identify the existing budget, resources, personnel, and potential for new public/private partnerships necessary for such a regulatory system; and

(D) consider risks identified by the actions undertaken to implement section 4 of this order.

(c) To promote the safe and responsible development and use of AI in the transportation sector, in consultation with relevant agencies:

(i) Within 30 days of the date of this order, the Secretary of Transportation shall direct the Nontraditional and Emerging Transportation Technology (NETT) Council to assess the need for information, technical assistance, and guidance regarding the use of AI in transportation. The Secretary of Transportation shall further direct the NETT Council, as part of any such efforts, to:

(A) support existing and future initiatives to pilot transportation-related applications of AI, as they align with policy priorities articulated in the Department of Transportation's (DOT) Innovation

Principles, including, as appropriate, through technical assistance and connecting stakeholders;

(B) evaluate the outcomes of such pilot programs in order to assess when DOT, or other Federal or State agencies, have sufficient information to take regulatory actions, as appropriate, and recommend appropriate actions when that information is available; and

(C) establish a new DOT Cross-Modal Executive Working Group, which will consist of members from different divisions of DOT and coordinate applicable work among these divisions, to solicit and use relevant input from appropriate stakeholders.

(ii) Within 90 days of the date of this order, the Secretary of Transportation shall direct appropriate Federal Advisory Committees of the DOT to provide advice on the safe and responsible use of AI in transportation. The committees shall include the Advanced Aviation Advisory Committee, the Transforming Transportation Advisory Committee, and the Intelligent Transportation Systems Program Advisory Committee.

(iii) Within 180 days of the date of this order, the Secretary of Transportation shall direct the Advanced Research Projects Agency-Infrastructure (ARPA-I) to explore the transportation-related opportunities and challenges of AI — including regarding software-defined AI enhancements impacting autonomous mobility ecosystems. The Secretary of Transportation shall further encourage ARPA-I to prioritize the allocation of grants to those opportunities, as appropriate. The work tasked to ARPA-I shall include soliciting input on these topics through a public consultation process, such as an RFI.

(d) To help ensure the responsible development and deployment of AI in the education sector, the Secretary of Education shall, within 365 days of the date of this order, develop resources, policies, and guidance regarding AI. These resources shall address safe, responsible, and non-

discriminatory uses of AI in education, including the impact AI systems have on vulnerable and underserved communities, and shall be developed in consultation with stakeholders as appropriate. They shall also include the development of an "AI toolkit" for education leaders implementing recommendations from the Department of Education's AI and the Future of Teaching and Learning report, including appropriate human review of AI decisions, designing AI systems to enhance trust and safety and align with privacy-related laws and regulations in the educational context, and developing education-specific guardrails.

(e) The Federal Communications Commission is encouraged to consider actions related to how AI will affect communications networks and consumers, including by:

(i) examining the potential for AI to improve spectrum management, increase the efficiency of non-Federal spectrum usage, and expand opportunities for the sharing of non-Federal spectrum;

(ii) coordinating with the National Telecommunications and Information Administration to create opportunities for sharing spectrum between Federal and non-Federal spectrum operations;

(iii) providing support for efforts to improve network security, resiliency, and interoperability using next-generation technologies that incorporate AI, including self-healing networks, 6G, and Open RAN; and

(iv) encouraging, including through rulemaking, efforts to combat unwanted robocalls and robotexts that are facilitated or exacerbated by AI and to deploy AI technologies that better serve consumers by blocking unwanted robocalls and robotexts.

22.2.9 Section 9. Protecting privacy

(a) To mitigate privacy risks potentially exacerbated by AI — including by AI's facilitation of the collection or use of information about

individuals, or the making of inferences about individuals — the Director of OMB shall:

(i) evaluate and take steps to identify commercially available information (CAI) procured by agencies, particularly CAI that contains personally identifiable information and including CAI procured from data brokers and CAI procured and processed indirectly through vendors, in appropriate agency inventory and reporting processes (other than when it is used for the purposes of national security);

(ii) evaluate, in consultation with the Federal Privacy Council and the Interagency Council on Statistical Policy, agency standards and procedures associated with the collection, processing, maintenance, use, sharing, dissemination, and disposition of CAI that contains personally identifiable information (other than when it is used for the purposes of national security) to inform potential guidance to agencies on ways to mitigate privacy and confidentiality risks from agencies' activities related to CAI;

(iii) within 180 days of the date of this order, in consultation with the Attorney General, the Assistant to the President for Economic Policy, and the Director of OSTP, issue an RFI to inform potential revisions to guidance to agencies on implementing the privacy provisions of the E-Government Act of 2002 (Public Law 107-347). The RFI shall seek feedback regarding how privacy impact assessments may be more effective at mitigating privacy risks, including those that are further exacerbated by AI; and

(iv) take such steps as are necessary and appropriate, consistent with applicable law, to support and advance the near-term actions and long-term strategy identified through the RFI process, including issuing new or updated guidance or RFIs or consulting other agencies or the Federal Privacy Council.

(b) Within 365 days of the date of this order, to better enable agencies to use PETs to safeguard Americans' privacy from the potential threats exacerbated by AI, the Secretary of Commerce, acting through the Director of NIST, shall create guidelines for agencies to evaluate the efficacy of differential-privacy-guarantee protections, including for AI. The guidelines shall, at a minimum, describe the significant factors that bear on differential-privacy safeguards and common risks to realizing differential privacy in practice.

(c) To advance research, development, and implementation related to PETs:

(i) Within 120 days of the date of this order, the Director of NSF, in collaboration with the Secretary of Energy, shall fund the creation of a Research Coordination Network (RCN) dedicated to advancing privacy research and, in particular, the development, deployment, and scaling of PETs. The RCN shall serve to enable privacy researchers to share information, coordinate and collaborate in research, and develop standards for the privacy-research community.

(ii) Within 240 days of the date of this order, the Director of NSF shall engage with agencies to identify ongoing work and potential opportunities to incorporate PETs into their operations. The Director of NSF shall, where feasible and appropriate, prioritize research — including efforts to translate research discoveries into practical applications — that encourage the adoption of leading-edge PETs solutions for agencies' use, including through research engagement through the RCN described in subsection (c)(i) of this section.

(iii) The Director of NSF shall use the results of the United States-United Kingdom PETs Prize Challenge to inform the approaches taken, and opportunities identified, for PETs research and adoption.

22.2.10 Section 10. Advancing federal government use of AI

10.1. Providing guidance for AI management.

(a) To coordinate the use of AI across the Federal Government, within 60 days of the date of this order and on an ongoing basis as necessary, the Director of OMB shall convene and chair an interagency council to coordinate the development and use of AI in agencies' programs and operations, other than the use of AI in national security systems. The Director of OSTP shall serve as Vice Chair for the interagency council. The interagency council's membership shall include, at minimum, the heads of the agencies identified in 31 U.S.C. 901(b), the Director of National Intelligence, and other agencies as identified by the Chair. Until agencies designate their permanent Chief AI Officers consistent with the guidance described in subsection 10.1(b) of this section, they shall be represented on the interagency council by an appropriate official at the Assistant Secretary level or equivalent, as determined by the head of each agency.

(b) To provide guidance on Federal Government use of AI, within 150 days of the date of this order and updated periodically thereafter, the Director of OMB, in coordination with the Director of OSTP, and in consultation with the interagency council established in subsection 10.1(a) of this section, shall issue guidance to agencies to strengthen the effective and appropriate use of AI, advance AI innovation, and manage risks from AI in the Federal Government. The Director of OMB's guidance shall specify, to the extent appropriate and consistent with applicable law:

(i) the requirement to designate at each agency within 60 days of the issuance of the guidance a Chief Artificial Intelligence Officer who shall hold primary responsibility in their agency, in coordination with other responsible officials, for coordinating their agency's use of AI, promoting AI innovation in their agency, managing risks from their agency's use of AI, and carrying out the responsibilities described in

section 8(c) of Executive Order 13960 of December 3, 2020 (Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government), and section 4(b) of Executive Order 14091;

(ii) the Chief Artificial Intelligence Officers' roles, responsibilities, seniority, position, and reporting structures;

(iii) for the agencies identified in 31 U.S.C. 901(b), the creation of internal Artificial Intelligence Governance Boards, or other appropriate mechanisms, at each agency within 60 days of the issuance of the guidance to coordinate and govern AI issues through relevant senior leaders from across the agency;

(iv) required minimum risk-management practices for Government uses of AI that impact people's rights or safety, including, where appropriate, the following practices derived from OSTP's Blueprint for an AI Bill of Rights and the NIST AI Risk Management Framework: conducting public consultation; assessing data quality; assessing and mitigating disparate impacts and algorithmic discrimination; providing notice of the use of AI; continuously monitoring and evaluating deployed AI; and granting human consideration and remedies for adverse decisions made using AI;

(v) specific Federal Government uses of AI that are presumed by default to impact rights or safety;

(vi) recommendations to agencies to reduce barriers to the responsible use of AI, including barriers related to information technology infrastructure, data, workforce, budgetary restrictions, and cybersecurity processes;

(vii) requirements that agencies identified in 31 U.S.C. 901(b) develop AI strategies and pursue high-impact AI use cases;

(viii) in consultation with the Secretary of Commerce, the Secretary of Homeland Security, and the heads of other appropriate agencies as

determined by the Director of OMB, recommendations to agencies regarding:

(A) external testing for AI, including AI red-teaming for generative AI, to be developed in coordination with the Cybersecurity and Infrastructure Security Agency;

(B) testing and safeguards against discriminatory, misleading, inflammatory, unsafe, or deceptive outputs, as well as against producing child sexual abuse material and against producing nonconsensual intimate imagery of real individuals (including intimate digital depictions of the body or body parts of an identifiable individual), for generative AI;

(C) reasonable steps to watermark or otherwise label output from generative AI;

(D) application of the mandatory minimum risk-management practices defined under subsection 10.1(b)(iv) of this section to procured AI;

(E) independent evaluation of vendors' claims concerning both the effectiveness and risk mitigation of their AI offerings;

(F) documentation and oversight of procured AI;

(G) maximizing the value to agencies when relying on contractors to use and enrich Federal Government data for the purposes of AI development and operation;

(H) provision of incentives for the continuous improvement of procured AI; and

(I) training on AI in accordance with the principles set out in this order and in other references related to AI listed herein; and

(ix) requirements for public reporting on compliance with this guidance.

(c) To track agencies' AI progress, within 60 days of the issuance of the guidance established in subsection 10.1(b) of this section and updated periodically thereafter, the Director of OMB shall develop a method for

agencies to track and assess their ability to adopt AI into their programs and operations, manage its risks, and comply with Federal policy on AI. This method should draw on existing related efforts as appropriate and should address, as appropriate and consistent with applicable law, the practices, processes, and capabilities necessary for responsible AI adoption, training, and governance across, at a minimum, the areas of information technology infrastructure, data, workforce, leadership, and risk management.

(d) To assist agencies in implementing the guidance to be established in subsection 10.1(b) of this section:

(i) within 90 days of the issuance of the guidance, the Secretary of Commerce, acting through the Director of NIST, and in coordination with the Director of OMB and the Director of OSTP, shall develop guidelines, tools, and practices to support implementation of the minimum risk-management practices described in subsection 10.1(b)(iv) of this section; and

(ii) within 180 days of the issuance of the guidance, the Director of OMB shall develop an initial means to ensure that agency contracts for the acquisition of AI systems and services align with the guidance described in subsection 10.1(b) of this section and advance the other aims identified in section 7224(d)(1) of the Advancing American AI Act (Public Law 117-263, div. G, title LXXII, subtitle B).

(e) To improve transparency for agencies' use of AI, the Director of OMB shall, on an annual basis, issue instructions to agencies for the collection, reporting, and publication of agency AI use cases, pursuant to section 7225(a) of the Advancing American AI Act. Through these instructions, the Director shall, as appropriate, expand agencies' reporting on how they are managing risks from their AI use cases and update or replace the guidance originally established in section 5 of Executive Order 13960.

(f) To advance the responsible and secure use of generative AI in the Federal Government:

(i) As generative AI products become widely available and common in online platforms, agencies are discouraged from imposing broad general bans or blocks on agency use of generative AI. Agencies should instead limit access, as necessary, to specific generative AI services based on specific risk assessments; establish guidelines and limitations on the appropriate use of generative AI; and, with appropriate safeguards in place, provide their personnel and programs with access to secure and reliable generative AI capabilities, at least for the purposes of experimentation and routine tasks that carry a low risk of impacting Americans' rights. To protect Federal Government information, agencies are also encouraged to employ risk-management practices, such as training their staff on proper use, protection, dissemination, and disposition of Federal information; negotiating appropriate terms of service with vendors; implementing measures designed to ensure compliance with record-keeping, cybersecurity, confidentiality, privacy, and data protection requirements; and deploying other measures to prevent misuse of Federal Government information in generative AI.

(ii) Within 90 days of the date of this order, the Administrator of General Services, in coordination with the Director of OMB, and in consultation with the Federal Secure Cloud Advisory Committee and other relevant agencies as the Administrator of General Services may deem appropriate, shall develop and issue a framework for prioritizing critical and emerging technologies offerings in the Federal Risk and Authorization Management Program authorization process, starting with generative AI offerings that have the primary purpose of providing large language model-based chat interfaces, code-generation and debugging tools, and associated application programming

interfaces, as well as prompt-based image generators. This framework shall apply for no less than 2 years from the date of its issuance. Agency Chief Information Officers, Chief Information Security Officers, and authorizing officials are also encouraged to prioritize generative AI and other critical and emerging technologies in granting authorities for agency operation of information technology systems and any other applicable release or oversight processes, using continuous authorizations and approvals wherever feasible.

(iii) Within 180 days of the date of this order, the Director of the Office of Personnel Management (OPM), in coordination with the Director of OMB, shall develop guidance on the use of generative AI for work by the Federal workforce.

(g) Within 30 days of the date of this order, to increase agency investment in AI, the Technology Modernization Board shall consider, as it deems appropriate and consistent with applicable law, prioritizing funding for AI projects for the Technology Modernization Fund for a period of at least 1 year. Agencies are encouraged to submit to the Technology Modernization Fund project funding proposals that include AI — and particularly generative AI — in service of mission delivery.

(h) Within 180 days of the date of this order, to facilitate agencies' access to commercial AI capabilities, the Administrator of General Services, in coordination with the Director of OMB, and in collaboration with the Secretary of Defense, the Secretary of Homeland Security, the Director of National Intelligence, the Administrator of the National Aeronautics and Space Administration, and the head of any other agency identified by the Administrator of General Services, shall take steps consistent with applicable law to facilitate access to Federal Government-wide acquisition solutions for specified types of AI services and products, such as through the creation of a resource guide or other tools to assist

the acquisition workforce. Specified types of AI capabilities shall include generative AI and specialized computing infrastructure.

(i) The initial means, instructions, and guidance issued pursuant to subsections 10.1(a)-(h) of this section shall not apply to AI when it is used as a component of a national security system, which shall be addressed by the proposed National Security Memorandum described in subsection 4.8 of this order.

10.2. Increasing AI talent in government.

(a) Within 45 days of the date of this order, to plan a national surge in AI talent in the Federal Government, the Director of OSTP and the Director of OMB, in consultation with the Assistant to the President for National Security Affairs, the Assistant to the President for Economic Policy, the Assistant to the President and Domestic Policy Advisor, and the Assistant to the President and Director of the Gender Policy Council, shall identify priority mission areas for increased Federal Government AI talent, the types of talent that are highest priority to recruit and develop to ensure adequate implementation of this order and use of relevant enforcement and regulatory authorities to address AI risks, and accelerated hiring pathways.

(b) Within 45 days of the date of this order, to coordinate rapid advances in the capacity of the Federal AI workforce, the Assistant to the President and Deputy Chief of Staff for Policy, in coordination with the Director of OSTP and the Director of OMB, and in consultation with the National Cyber Director, shall convene an AI and Technology Talent Task Force, which shall include the Director of OPM, the Director of the General Services Administration's Technology Transformation Services, a representative from the Chief Human Capital Officers Council, the Assistant to the President for Presidential Personnel, members of appropriate agency technology talent programs, a representative of the Chief Data Officer Council, and a representative of the interagency council convened

under subsection 10.1(a) of this section. The Task Force's purpose shall be to accelerate and track the hiring of AI and AI-enabling talent across the Federal Government, including through the following actions:

(i) within 180 days of the date of this order, tracking and reporting progress to the President on increasing AI capacity across the Federal Government, including submitting to the President a report and recommendations for further increasing capacity;

(ii) identifying and circulating best practices for agencies to attract, hire, retain, train, and empower AI talent, including diversity, inclusion, and accessibility best practices, as well as to plan and budget adequately for AI workforce needs;

(iii) coordinating, in consultation with the Director of OPM, the use of fellowship programs and agency technology-talent programs and human-capital teams to build hiring capabilities, execute hires, and place AI talent to fill staffing gaps; and

(iv) convening a cross-agency forum for ongoing collaboration between AI professionals to share best practices and improve retention.

(c) Within 45 days of the date of this order, to advance existing Federal technology talent programs, the United States Digital Service, Presidential Innovation Fellowship, United States Digital Corps, OPM, and technology talent programs at agencies, with support from the AI and Technology Talent Task Force described in subsection 10.2(b) of this section, as appropriate and permitted by law, shall develop and begin to implement plans to support the rapid recruitment of individuals as part of a Federal Government-wide AI talent surge to accelerate the placement of key AI and AI-enabling talent in high-priority areas and to advance agencies' data and technology strategies.

(d) To meet the critical hiring need for qualified personnel to execute the initiatives in this order, and to improve Federal hiring practices for AI

talent, the Director of OPM, in consultation with the Director of OMB, shall:

(i) within 60 days of the date of this order, conduct an evidence-based review on the need for hiring and workplace flexibility, including Federal Government-wide direct-hire authority for AI and related data-science and technical roles, and, where the Director of OPM finds such authority is appropriate, grant it; this review shall include the following job series at all General Schedule (GS) levels: IT Specialist (2210), Computer Scientist (1550), Computer Engineer (0854), and Program Analyst (0343) focused on AI, and any subsequently developed job series derived from these job series;

(ii) within 60 days of the date of this order, consider authorizing the use of excepted service appointments under 5 C.F.R. 213.3102(i)(3) to address the need for hiring additional staff to implement directives of this order;

(iii) within 90 days of the date of this order, coordinate a pooled-hiring action informed by subject-matter experts and using skills-based assessments to support the recruitment of AI talent across agencies;

(iv) within 120 days of the date of this order, as appropriate and permitted by law, issue guidance for agency application of existing pay flexibilities or incentive pay programs for AI, AI-enabling, and other key technical positions to facilitate appropriate use of current pay incentives;

(v) within 180 days of the date of this order, establish guidance and policy on skills-based, Federal Government-wide hiring of AI, data, and technology talent in order to increase access to those with nontraditional academic backgrounds to Federal AI, data, and technology roles;

(vi) within 180 days of the date of this order, establish an interagency working group, staffed with both human-resources professionals and

recruiting technical experts, to facilitate Federal Government-wide hiring of people with AI and other technical skills;

(vii) within 180 days of the date of this order, review existing Executive Core Qualifications (ECQs) for Senior Executive Service (SES) positions informed by data and AI literacy competencies and, within 365 days of the date of this order, implement new ECQs as appropriate in the SES assessment process;

(viii) within 180 days of the date of this order, complete a review of competencies for civil engineers (GS-0810 series) and, if applicable, other related occupations, and make recommendations for ensuring that adequate AI expertise and credentials in these occupations in the Federal Government reflect the increased use of AI in critical infrastructure; and

(ix) work with the Security, Suitability, and Credentialing Performance Accountability Council to assess mechanisms to streamline and accelerate personnel-vetting requirements, as appropriate, to support AI and fields related to other critical and emerging technologies.

(e) To expand the use of special authorities for AI hiring and retention, agencies shall use all appropriate hiring authorities, including Schedule A(r) excepted service hiring and direct-hire authority, as applicable and appropriate, to hire AI talent and AI-enabling talent rapidly. In addition to participating in OPM-led pooled hiring actions, agencies shall collaborate, where appropriate, on agency-led pooled hiring under the Competitive Service Act of 2015 (Public Law 114-137) and other shared hiring. Agencies shall also, where applicable, use existing incentives, pay-setting authorities, and other compensation flexibilities, similar to those used for cyber and information technology positions, for AI and data-science professionals, as well as plain-language job titles, to help recruit and retain these highly skilled professionals. Agencies shall ensure that AI and other related talent needs (such as technology

governance and privacy) are reflected in strategic workforce planning and budget formulation.

(f) To facilitate the hiring of data scientists, the Chief Data Officer Council shall develop a position-description library for data scientists (job series 1560) and a hiring guide to support agencies in hiring data scientists.

(g) To help train the Federal workforce on AI issues, the head of each agency shall implement — or increase the availability and use of — AI training and familiarization programs for employees, managers, and leadership in technology as well as relevant policy, managerial, procurement, regulatory, ethical, governance, and legal fields. Such training programs should, for example, empower Federal employees, managers, and leaders to develop and maintain an operating knowledge of emerging AI technologies to assess opportunities to use these technologies to enhance the delivery of services to the public, and to mitigate risks associated with these technologies. Agencies that provide professional-development opportunities, grants, or funds for their staff should take appropriate steps to ensure that employees who do not serve in traditional technical roles, such as policy, managerial, procurement, or legal fields, are nonetheless eligible to receive funding for programs and courses that focus on AI, machine learning, data science, or other related subject areas.

(h) Within 180 days of the date of this order, to address gaps in AI talent for national defense, the Secretary of Defense shall submit a report to the President through the Assistant to the President for National Security Affairs that includes:

(i) recommendations to address challenges in the Department of Defense's ability to hire certain noncitizens, including at the Science and Technology Reinvention Laboratories;

(ii) recommendations to clarify and streamline processes for accessing classified information for certain noncitizens through Limited Access Authorization at Department of Defense laboratories;

(iii) recommendations for the appropriate use of enlistment authority under 10 U.S.C. 504(b)(2) for experts in AI and other critical and emerging technologies; and

(iv) recommendations for the Department of Defense and the Department of Homeland Security to work together to enhance the use of appropriate authorities for the retention of certain noncitizens of vital importance to national security by the Department of Defense and the Department of Homeland Security.

22.2.11 Section 11. Strengthening American leadership abroad

(a) To strengthen United States leadership of global efforts to unlock AI's potential and meet its challenges, the Secretary of State, in coordination with the Assistant to the President for National Security Affairs, the Assistant to the President for Economic Policy, the Director of OSTP, and the heads of other relevant agencies as appropriate, shall:

(i) lead efforts outside of military and intelligence areas to expand engagements with international allies and partners in relevant bilateral, multilateral, and multi-stakeholder fora to advance those allies' and partners' understanding of existing and planned AI-related guidance and policies of the United States, as well as to enhance international collaboration; and

(ii) lead efforts to establish a strong international framework for managing the risks and harnessing the benefits of AI, including by encouraging international allies and partners to support voluntary commitments similar to those that United States companies have made in pursuit of these objectives and coordinating the activities directed by subsections (b), (c), (d), and (e) of this section, and to develop

common regulatory and other accountability principles for foreign nations, including to manage the risk that AI systems pose.

(b) To advance responsible global technical standards for AI development and use outside of military and intelligence areas, the Secretary of Commerce, in coordination with the Secretary of State and the heads of other relevant agencies as appropriate, shall lead preparations for a coordinated effort with key international allies and partners and with standards development organizations, to drive the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing. In particular, the Secretary of Commerce shall:

(i) within 270 days of the date of this order, establish a plan for global engagement on promoting and developing AI standards, with lines of effort that may include:

(A) AI nomenclature and terminology;

(B) best practices regarding data capture, processing, protection, privacy, confidentiality, handling, and analysis;

(C) trustworthiness, verification, and assurance of AI systems; and

(D) AI risk management;

(ii) within 180 days of the date the plan is established, submit a report to the President on priority actions taken pursuant to the plan; and

(iii) ensure that such efforts are guided by principles set out in the NIST AI Risk Management Framework and United States Government National Standards Strategy for Critical and Emerging Technology.

(c) Within 365 days of the date of this order, to promote safe, responsible, and rights-affirming development and deployment of AI abroad:

(i) The Secretary of State and the Administrator of the United States Agency for International Development, in coordination with the

Secretary of Commerce, acting through the director of NIST, shall publish an AI in Global Development Playbook that incorporates the AI Risk Management Framework's principles, guidelines, and best practices into the social, technical, economic, governance, human rights, and security conditions of contexts beyond United States borders. As part of this work, the Secretary of State and the Administrator of the United States Agency for International Development shall draw on lessons learned from programmatic uses of AI in global development.

(ii) The Secretary of State and the Administrator of the United States Agency for International Development, in collaboration with the Secretary of Energy and the Director of NSF, shall develop a Global AI Research Agenda to guide the objectives and implementation of AI-related research in contexts beyond United States borders. The Agenda shall:

(A) include principles, guidelines, priorities, and best practices aimed at ensuring the safe, responsible, beneficial, and sustainable global development and adoption of AI; and

(B) address AI's labor-market implications across international contexts, including by recommending risk mitigations.

(d) To address cross-border and global AI risks to critical infrastructure, the Secretary of Homeland Security, in coordination with the Secretary of State, and in consultation with the heads of other relevant agencies as the Secretary of Homeland Security deems appropriate, shall lead efforts with international allies and partners to enhance cooperation to prevent, respond to, and recover from potential critical infrastructure disruptions resulting from incorporation of AI into critical infrastructure systems or malicious use of AI.

(i) Within 270 days of the date of this order, the Secretary of Homeland Security, in coordination with the Secretary of State, shall

develop a plan for multilateral engagements to encourage the adoption of the AI safety and security guidelines for use by critical infrastructure owners and operators developed in section 4.3(a) of this order.

(ii) Within 180 days of establishing the plan described in subsection (d)(i) of this section, the Secretary of Homeland Security shall submit a report to the President on priority actions to mitigate cross-border risks to critical United States infrastructure.

22.2.12 Section 12. Implementation

(a) There is established, within the Executive Office of the President, the White House Artificial Intelligence Council (White House AI Council). The function of the White House AI Council is to coordinate the activities of agencies across the Federal Government to ensure the effective formulation, development, communication, industry engagement related to, and timely implementation of AI-related policies, including policies set forth in this order.

(b) The Assistant to the President and Deputy Chief of Staff for Policy shall serve as Chair of the White House AI Council.

(c) In addition to the Chair, the White House AI Council shall consist of the following members, or their designees:

(i) the Secretary of State;

(ii) the Secretary of the Treasury;

(iii) the Secretary of Defense;

(iv) the Attorney General;

(v) the Secretary of Agriculture;

(vi) the Secretary of Commerce;

(vii) the Secretary of Labor;

(viii) the Secretary of HHS;

(ix) the Secretary of Housing and Urban Development;

(x) the Secretary of Transportation;

(xi) the Secretary of Energy;

(xii) the Secretary of Education;

(xiii) the Secretary of Veterans Affairs;

(xiv) the Secretary of Homeland Security;

(xv) the Administrator of the Small Business Administration;

(xvi) the Administrator of the United States Agency for International Development;

(xvii) the Director of National Intelligence;

(xviii) the Director of NSF;

(xix) the Director of OMB;

(xx) the Director of OSTP;

(xxi) the Assistant to the President for National Security Affairs;

(xxii) the Assistant to the President for Economic Policy;

(xxiii) the Assistant to the President and Domestic Policy Advisor;

(xxiv) the Assistant to the President and Chief of Staff to the Vice President;

(xxv) the Assistant to the President and Director of the Gender Policy Council;

(xxvi) the Chairman of the Council of Economic Advisers;

(xxvii) the National Cyber Director;

(xxviii) the Chairman of the Joint Chiefs of Staff; and

(xxix) the heads of such other agencies, independent regulatory agencies, and executive offices as the Chair may from time to time designate or invite to participate.

(d) The Chair may create and coordinate subgroups consisting of White House AI Council members or their designees, as appropriate.

Sec. 13. General provisions

(a) Nothing in this order shall be construed to impair or otherwise affect:

(i) the authority granted by law to an executive department or agency, or the head thereof; or

(ii) the functions of the Director of the Office of Management and Budget relating to budgetary, administrative, or legislative proposals.

(b) This order shall be implemented consistent with applicable law and subject to the availability of appropriations.

(c) This order is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

JOSEPH R. BIDEN JR.

THE WHITE HOUSE

October 30, 2023

PART IV

PRIVATE SECTOR AI GUIDELEINES, STANDARDS REGULATIONS AND ETHICS: GLOBAL

Private global media companies play a key role in development, deployment and user control of AI programmes. They all developed in last few years internal guidelines and principles for AI development and use. They do it voluntarily, but also under the pressure of binding legal standards, decided by governments (above Part III). The 2x4 giants are Google, Apple, Facebook/Meta and Amazon (GAFA) in the US and Baidu, Alibaba, Tencent and Huawei (BATH) in China. The AI Principles of five of the eight giants are presented below (chapters 23-27).

23

MICROSOFT RESPONSIBLE AI STANDARD

23.1 Introduction

The "Microsoft Responsible AI Standard v2" (Version 2)[434] was published in June 2022. It includes six categories of goals: Accountability Goals, Transparency Goals, Fairness Goals, Reliability and Safety Goals, Privacy and Security Goals, Inclusiveness Goals. The following text[435] is a summary explanation of the extended Standard.

Microsoft offers online manifold materials on its Responsible AI[436] commitment: guidelines, implementation procedures, academic reference guide[437] etc. Microsoft also published in May 2024 for the first time its "Responsible AI Transparency Report."[438] The business model is

---

[434] https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf.

[435] Microsoft, *What is Responsible AI*? Article published on MS Website on 23 Sept 2024. https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai?view=azureml-api-2#:~:text=Microsoft%20developed%20a%20Responsible%20AI%20Standard.%20It%27s%20a,safety%2C%20privacy%20and%20security%2C%20inclusiveness%2C%20transparency%2C%20and%20accountability.

[436] https://www.microsoft.com/en-us/ai/responsible-ai

[437] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5dlCb

[438] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1l5BO

clear: Microsoft offers AI solutions for individuals and companies and needs to be trusted.

The Editors

23.2 Microsoft: What is responsible AI?

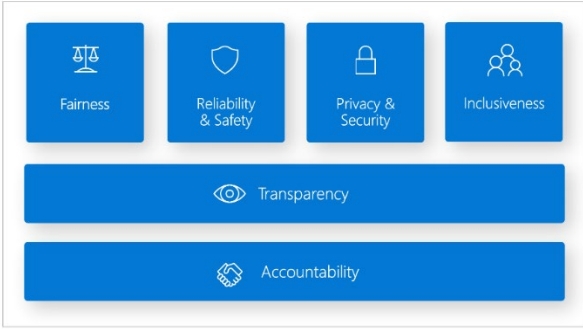
Responsible Artificial Intelligence (Responsible AI) is an approach to developing, assessing, and deploying AI systems in a safe, trustworthy, and ethical way. AI systems are the product of many decisions made by those who develop and deploy them. From system purpose to how people interact with AI systems, Responsible AI can help proactively guide these decisions toward more beneficial and equitable outcomes. That means keeping people and their goals at the center of system design decisions and respecting enduring values like fairness, reliability, and transparency.

Microsoft developed a Responsible AI Standard[439]. It's a framework for building AI systems according to six principles: fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. For Microsoft, these principles are the cornerstone of a responsible and trustworthy approach to AI, especially as intelligent technology becomes more prevalent in products and services that people use every day.

This article demonstrates how Azure Machine Learning supports tools for enabling developers and data scientists to implement and operationalize the six principles.

---

439 https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf



Fairness and inclusiveness: AI systems should treat everyone fairly and avoid affecting similarly situated groups of people in different ways. For example, when AI systems provide guidance on medical treatment, loan applications, or employment, they should make the same recommendations to everyone who has similar symptoms, financial circumstances, or professional qualifications.

Fairness and inclusiveness in Azure Machine Learning: The fairness assessment[440] component of the Responsible AI dashboard[441] enables data scientists and developers to assess model fairness across sensitive groups defined in terms of gender, ethnicity, age, and other characteristics.

Reliability and safety: To build trust, it's critical that AI systems operate reliably, safely, and consistently. These systems should be able to operate as they were originally designed, respond safely to unanticipated conditions, and resist harmful manipulation. How they behave and the variety of conditions they can handle reflect the range of situations and circumstances that developers anticipated during design and testing.

---

[440] https://learn.microsoft.com/en-us/azure/machine-learning/concept-fairness-ml?view=azureml-api-2

[441] https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard?view=azureml-api-2

Reliability and safety in Azure Machine Learning: The error analysis[442] component of the Responsible AI dashboard[443] enables data scientists and developers to:

- Get a deep understanding of how failure is distributed for a model.
- Identify cohorts (subsets) of data with a higher error rate than the overall benchmark.

These discrepancies might occur when the system or model underperforms for specific demographic groups or for infrequently observed input conditions in the training data.

Transparency: When AI systems help inform decisions that have tremendous impacts on people's lives, it's critical that people understand how those decisions were made. For example, a bank might use an AI system to decide whether a person is creditworthy. A company might use an AI system to determine the most qualified candidates to hire.

A crucial part of transparency is interpretability: the useful explanation of the behavior of AI systems and their components. Improving interpretability requires stakeholders to comprehend how and why AI systems function the way they do. The stakeholders can then identify potential performance issues, fairness issues, exclusionary practices, or unintended outcomes.

Transparency in Azure Machine Learning: The model interpretability[444] and counterfactual what-if[445] components of the Responsible AI

---

[442] https://learn.microsoft.com/en-us/azure/machine-learning/concept-error-analysis?view=azureml-api-2

[443] https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard?view=azureml-api-2

[444] https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability?view=azureml-api-2

[445] https://learn.microsoft.com/en-us/azure/machine-learning/concept-counterfactual-analysis?view=azureml-api-2

dashboard[446] enable data scientists and developers to generate human-understandable descriptions of the predictions of a model.

The model interpretability component provides multiple views into a model's behaviour:

- Global explanations. For example, what features affect the overall behaviour of a loan allocation model?
- Local explanations. For example, why was a customer's loan application approved or rejected?
- Model explanations for a selected cohort of data points. For example, what features affect the overall behaviour of a loan allocation model for low-income applicants?

The counterfactual what-if component enables understanding and debugging a machine learning model in terms of how it reacts to feature changes and perturbations.

Azure Machine Learning also supports a Responsible AI scorecard.[447] The scorecard is a customizable PDF report that developers can easily configure, generate, download, and share with their technical and non-technical stakeholders to educate them about their datasets and models health, achieve compliance, and build trust. This scorecard can also be used in audit reviews to uncover the characteristics of machine learning models.

Privacy and security: As AI becomes more prevalent, protecting privacy and securing personal and business information are becoming more important and complex. With AI, privacy and data security require close attention because access to data is essential for AI systems to make

---

446 https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard?view=azureml-api-2

447 https://learn.microsoft.com/en-us/azure/machine-learning/how-to-responsible-ai-scorecard?view=azureml-api-2

accurate and informed predictions and decisions about people. AI systems must comply with privacy laws that:

- Require transparency about the collection, use, and storage of data.
- Mandate that consumers have appropriate controls to choose how their data is used.

Privacy and security in Azure Machine Learning: Azure Machine Learning enables administrators and developers to create a secure configuration that complies[448] with their companies' policies. With Azure Machine Learning and the Azure platform, users can:

- Restrict access to resources and operations by user account or group.
- Restrict incoming and outgoing network communications.
- Encrypt data in transit and at rest.
- Scan for vulnerabilities.
- Apply and audit configuration policies.

Microsoft also created two open-source packages that can enable further implementation of privacy and security principles:

- SmartNoise:[449] Differential privacy is a set of systems and practices that help keep the data of individuals safe and private. In machine learning solutions, differential privacy might be required for regulatory compliance. SmartNoise is an open-source project (co-developed by Microsoft) that contains components for building differentially private systems that are global.

[448] https://learn.microsoft.com/en-us/azure/machine-learning/concept-enterprise-security?view=azureml-api-2

[449] https://github.com/opendifferentialprivacy/smartnoise-core

- Counterfit:[450] Counterfit is an open-source project that comprises a command-line tool and generic automation layer to allow developers to simulate cyberattacks against AI systems. Anyone can download the tool and deploy it through Azure Cloud Shell to run in a browser, or deploy it locally in an Anaconda Python environment. It can assess AI models hosted in various cloud environments, on-premises, or in the edge. The tool is agnostic to AI models and supports various data types, including text, images, or generic input.

Accountability: The people who design and deploy AI systems must be accountable for how their systems operate. Organizations should draw upon industry standards to develop accountability norms. These norms can ensure that AI systems aren't the final authority on any decision that affects people's lives. They can also ensure that humans maintain meaningful control over otherwise highly autonomous AI systems.

Accountability in Azure Machine Learning: Machine learning operations (MLOps)[451] is based on DevOps principles and practices that increase the efficiency of AI workflows. Azure Machine Learning provides the following MLOps capabilities for better accountability of your AI systems:

- Register, package, and deploy models from anywhere. You can also track the associated metadata that's required to use the model.
- Capture the governance data for the end-to-end machine learning lifecycle. The logged lineage information can

[450] https://github.com/Azure/counterfit/

[451] https://learn.microsoft.com/en-us/azure/machine-learning/concept-model-management-and-deployment?view=azureml-api-2

include who is publishing models, why changes were made, and when models were deployed or used in production.

- Notify and alert on events in the machine learning lifecycle. Examples include experiment completion, model registration, model deployment, and data drift detection.
- Monitor applications for operational issues and issues related to machine learning. Compare model inputs between training and inference, explore model-specific metrics, and provide monitoring and alerts on your machine learning infrastructure.

Besides the MLOps capabilities, the Responsible AI scorecard[452] in Azure Machine Learning creates accountability by enabling cross-stakeholder communications. The scorecard also creates accountability by empowering developers to configure, download, and share their model health insights with their technical and non-technical stakeholders about AI data and model health. Sharing these insights can help build trust.

The machine learning platform also enables decision-making by informing business decisions through:

- Data-driven insights, to help stakeholders understand causal treatment effects on an outcome, by using historical data only. For example, "How would a medicine affect a patient's blood pressure?" These insights are provided through the causal inference[453] component of the Responsible AI dashboard.[454]

---

[452] https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-scorecard?view=azureml-api-2

[453] https://learn.microsoft.com/en-us/azure/machine-learning/concept-causal-inference?view=azureml-api-2

[454] https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard?view=azureml-api-2

- Model-driven insights, to answer users' questions (such as "What can I do to get a different outcome from your AI next time?") so they can take action. Such insights are provided to data scientists through the counterfactual what-if[455] component of the Responsible AI dashboard.[456]

Next steps

- For more information on how to implement Responsible AI in Azure Machine Learning, see Responsible AI dashboard.[457]
- Learn how to generate the Responsible AI dashboard via CLI and SDK[458] or Azure Machine Learning studio UI.[459]
- Learn how to generate a Responsible AI scorecard[460] based on the insights observed in your Responsible AI dashboard.
- Learn about the Responsible AI Standard[461] for building AI systems according to six key principles.

---

455 https://learn.microsoft.com/en-us/azure/machine-learning/concept-counterfactual-analysis?view=azureml-api-2

456 https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard?view=azureml-api-2

457 https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard?view=azureml-api-2

458 https://learn.microsoft.com/en-us/azure/machine-learning/how-to-responsible-ai-dashboard-sdk-cli?view=azureml-api-2

459 https://learn.microsoft.com/en-us/azure/machine-learning/how-to-responsible-ai-dashboard-ui?view=azureml-api-2

460 https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-scorecard?view=azureml-api-2

461 https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

24

GOOGLE AI PRINCIPLES

24.1 Introduction

Google like the other global media giants developed in last few years internal guidelines and principles for AI development and use. They do it voluntarily, but also under the pressure of legally binding standards and legislation of governments. The following text are the "Google AI Principles", republished as Preface of the 2023 update[462].

The "Ethics of Advanced AI Assistance", published in April 2024 offers ethical reflection and support for individual and collective AI users.[463] It is a detailed guide and handbook about values, risks of AI and how to use it responsively as assistant to humans.

The Editors

---

[462] Google, *AI Principles Progress Update 2023*, Preface, 3-5. https://publicpolicy.google/responsible-ai/.

[463] Google, *Ethics of Advanced AI Assistance.* Google Deep Mind, 28 April 2024. https://arxiv.org/pdf/2404.16244.

24.2 Google AI principles: Objectives for AI applications

1. Be socially beneficial. The expanded reach of new technologies increasingly touches society as a whole. Advances in AI will have transformative impacts in a wide range of fields, including healthcare, security, energy, transportation, manufacturing, and entertainment. As we consider potential development and use of AI technologies, we will take into account a broad range of social and economic factors, and will proceed where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides.

AI also enhances our ability to understand the meaning of content at scale. We will strive to make high-quality and accurate information readily available using AI, while continuing to respect cultural, social, and legal norms in the countries or regions where we operate. And we will continue to thoughtfully evaluate when to make our technologies available on a non-commercial basis.

2. Avoid creating or reinforcing unfair bias. AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. We recognize that distinguishing fair from unfair biases is not always simple, and differs across cultures and societies. We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.

3. Be built & tested for safety. We will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm. We will design our AI systems to be appropriately cautious, and seek to develop them in accordance with best practices in AI safety research. In appropriate cases, we will test AI technologies in constrained environments and monitor their operation after deployment.

4. Be accountable to people. We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal.

Our AI technologies will be subject to appropriate human direction and control.

5. Incorporate privacy design principles. We will incorporate our privacy principles in the development and use of our AI technologies. We will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data.

6. Uphold high standards of scientific excellence. Technological innovation is rooted in the scientific method and a commitment to open inquiry, intellectual rigor, integrity, and collaboration. AI tools have the potential to unlock new realms of scientific research and knowledge in critical domains like biology, chemistry, medicine, and environmental sciences. We aspire to high standards of scientific excellence as we work to progress AI development.

We will work with a range of stakeholders to promote thoughtful leadership in this area, drawing on scientifically rigorous and multidisciplinary approaches. And we will responsibly share AI knowledge by publishing educational materials, best practices, and research that enable more people to develop useful AI applications.

7. Be made available for uses that accord with these principles. Many technologies have multiple uses. We will work to limit potentially harmful or abusive applications. As we develop and deploy AI technologies, we will evaluate likely uses in light of the following factors:

- Primary purpose and use: the primary purpose and likely use of a technology and application, including how closely the solution is related to or adaptable to a harmful use
- Nature and uniqueness: whether we are making available technology that is unique or more generally available
- Scale: whether the use of this technology will have significant impact

- Nature of Google's involvement: whether we are providing general-purpose tools, integrating tools for customers, or developing custom solutions.

AI applications we will not pursue

In addition to the above objectives, we will not design or deploy AI in the following application areas:

1. Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks and will incorporate appropriate safety constraints.
2. Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
3. Technologies that gather or use information for surveillance violating internationally accepted norms.
4. Technologies whose purpose contravenes widely accepted principles of international law and human rights

As our experience in this space deepens, this list may evolve.

24.3 The ethics of advanced AI assistance

What is an AI Assistant? Google defines it in its 273-pages guide as follows: "We define an AI assistant here as an artificial agent with a natural language interface, the function of which is to plan and execute sequences of actions on the user's behalf across one or more domains and in line with the user's expectations."[464] The text on AI Assistant is a detailed guide and handbook for AI users about values, risks of AI and how to use it responsively as assistant to humans. Standards and legislations

[464] Google, *The Ethics of Advanced AI Assistance.* Google Deep Mind, 2024, 15. https://arxiv.org/pdf/2404.16244,

as in Part III and IV of this book on AI Governance Ethics are important frames, but need the concrete guidance as this google document. The following text is the detailed table of content which may inspire the questions and further curiosity to consult the full text.[465]

[465] Google, *The Ethics of Advanced AI Assistance.* Google Deep Mind, 2024. Free download https://arxiv.org/pdf/2404.16244.

25

SINGULARITYNET DECENTRALISED AI MARKETPLACE

25.1 Introduction

"SingularityNET[466] is the world's leading decentralised AI marketplace, running on blockchain. Our core mission is the development of Artificial General Intelligence (AGI) for a beneficial technological Singularity".[467] "An 'AGI' that is not dependent on any central entity, that is open for anyone and not restricted to the narrow goals of a single corporation or even a single country." The specific character compared to listed private companies (in this book chapters 23, 24, 26, 27) is the democratic governance of SingularityNET. The SingularityNET-Ambassadors-Program includes an AI Ethics workgroup[468] to further develop democratically ethics aspects. The following text is an excerpt from the "SingularityNET White Paper 2.0".[469]

The Editors

---

[466] https://singularitynet.io/.

[467] https://app.singularitynetdapps.com/.

[468] https://github.com/SingularityNet-Ambassador-Program/AI-Ethics-Workgroup.

[469] SingularityNET White Paper 2.0, 2019, 6 and 34-37.

25.2 SingularityNET Vision

Inspiration

The inevitability of a technological singularity is increasingly accepted throughout the technology and business worlds. Knowledgeable people are realizing that the next few decades will see a transition to a new society and economy in which machine intelligence is the dominant factor. For this to occur, swarms of machine and organic intelligences must network together to produce emergent "global brain" dynamics of unprecedented complexity and sophistication with power and flexibility none alone would have.[470]

Markets display elements of cognitive synergy—agents in an economy each pursue their own relatively simple goals, but patterns with higher-order goals emerge from their interactions. SingularityNET is not only a collection of AIs, it is a market. It is designed to harness self-organizing swarm intelligence to create a whole greater than the sum of its parts.

Blockchain allows us to program economic rules in a digital environment and AI software to seamlessly interact with them. The path to creating a positive "global brain" is challenging. A technological singularity could have unprecedented benefits but also poses unprecedented risk. The popular press is full of dire warnings about the dangers of artificial general intelligence. Among the challenges is the current set of protocols for collective action; in many respects, today's financial mechanisms and institutions would give us a risky ride to the singularity.

---

[470] Damien Broderick, *The Spike* (Tor Books, 1997); Ray Kurzweil, *The Singularity Is Near* (2006); Vernor Vinge, "The Coming Technological Singularity," *Whole Earth Review* 81 (1993): 88–95; Ben Goertzel, "Human-level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's 'The Singularity Is Near,' and McDermott's Critique of Kurzweil," *Artificial Intelligence* 171, no. 18 (2007), 1161–73.

New, more flexible, open, and rapidly adaptive economic structures and dynamics are needed.[471]

Blockchain, with its natively digital money, is a powerful tool for managing transactions in an economy dominated by machine intelligence. [472] However, blockchain is just a tool; there are important decisions to be made about how to use it. SingularityNET is a blockchain-based framework designed to serve the needs of AI agents as they interact with each other and with external customers. At its core, SingularityNET is a set of smart contract templates that AI agents can use to request that AI work be done, to exchange data, and to supply the results of AI work.

This framework is a network that meshes disparate elements into a collective intelligence, much like the different areas of the brain - each with its own speciality - mesh together. It is critical that this network be designed with positive principles in mind:

- Democratic governance on specific issues - if the community governs the system, then the system will tend to act for the benefit of the community;
- Encouragement that prompts innovative new agents to enter the network, and creation of the conditions necessary for agents to act in a manner that feeds the collective intelligence;
- Direction of a significant percentage of the network’s efforts toward causes of broad benefit;

SingularityNET has been designed to meet these requirements by

---

[471] Ben Goertzel, Ted Goertzel, and Zarathustra Goertzel, “The Global Brain and the Emerging Economy of Abundance: Mutualism, Open Collaboration, Exchange Networks and the Automated Commons,” *Technological Forecasting and Social Change* 114C (2016): 65–73. 2016, http://goertzel.org/OpenCollaboration.pdf.

[472] John Clippinger and David Bollier, *From Bitcoin to Burning Man and Beyond* (Off the Common Books, 2014).

- delivering intelligence services to corporations, other organizations, and individuals;
- fostering the emergence of increasingly powerful distributed general intelligence; and
- deploying artificial intelligence for ever-increasing benefit to as many humans and other sentient beings as possible.

SingularityNET is designed both to be highly valuable now and to lay the groundwork for the emergence of a self-modifying, decentralized "artificial cognitive organism" with the eventual potential for general intelligence and beneficial ethical characteristics beyond the human level. It is a practical design inspired by long theoretical thinking and prototyping by our founders regarding artificial general intelligence[473], open-ended intelligence[474], the global brain[475], and more.

25.3 SingularityNET democratic governance

SingularityNET is not only a network of AI agents but also a network of humans who use, create, rate, and otherwise interact with the AIs. Because it is a decentralized organization, the ongoing health and growth of SingularityNET will rely on democratic decision-making by network participants. A democratic process is used to make decisions regarding network operation and to allocate newly minted AGI tokens.

---

[473] Ben Goertzel, *The AGI Revolution* (Humanity+ Press, 2016).

[474] David Weaver Weinbaum and Viktoras Veitas, "Open-ended Intelligence," *in International Conference on Artificial General Intelligence* (Springer, 2016): 43–52, https://arxiv.org/abs/1505.06366.

[475] 6 F. Heylighen, "The Global Superorganism: "An Evolutionary Cybernetic Model of the Emerging Network Society," *Social Evolution and History* 6, no. 1 (2007).

25.3.1 Reputation and stake-based voting

Voting is filtered by reputation; only Agents with a base reputation above the threshold of 2 will be counted. Furthermore, only Agents whose owners have been verified by appropriate KYC procedures will be permitted to vote (although other Agents can still participate in the network by offering or purchasing services).

The initial default plan is to use standard KYC methodology, likely via partnership with an external firm specializing in KYC for blockchain-based enterprises. Before year 4 of the network's operation, this will be replaced by a decentralized KYC methodology through which Agents are "KYC'd" by other Agents rather than any central authority. One possible approach is essentially a "verification federation" consisting of Agents that are democratically approved to perform KYC functions.

The amount of voting power that an owner (a verified entity that owns an agent) has regarding core network operation issues and the distribution of the future development reserve is given by the following formula:

Let stake (O) denote the total stake of owner O across all its Agents (i.e. its total amount of AGI token holdings); let stake (A) denote the stake of a particular Agent A; and let rep(A) denote the base reputation of Agent A. Let ag(O) denote the set of Agents owned by owner O.

We use the following definition:

$$\Psi(x) = \begin{cases} c * x \text{ if } x < L \\ c * (L + log_2(x - L + 2)) \text{ if } x \geq L \end{cases}$$

Here L is a boundary so that for x < L, the function Ψ behaves piecewise linearly and for x $\geqslant$ L, the function Ψ behaves logarithmically (and c is just an arbitrary normalizing factor). Then we set

$$\text{Votes} = log_2(\text{stake}(O)) * \sum_{A \in \text{ag}(O)} \Psi(\text{stake}(A))(\text{rep}(A) - 1)$$

The combination of reputation and stake in this formula gives more voting power to more highly rated entities while preventing attacks involving large numbers of sock puppet agents that have good reputations but carry out few transactions. At a high level, one can think of this voting formula as a sort of "Proof of Contribution":

- The use of the logarithmic function in the first term of the formula means that owners with more AGI tokens get to vote more, but that once the amount of their token ownership exceeds L, their voting power increases according to the order of magnitude of their AGI ownership rather than linearly.
- The second term in the formula means that owners whose agents are doing useful (highly reputable) things get more voting power. The use of the Ψ function is intended to avoid a dynamic in which owners are rewarded for splitting up their AI functions among many small agents, each with a tiny stake but a good reputation. For stakes up to size L, there is no reward for splitting up agents smaller than that size. For stakes above L, there is some reward for splitting up agents smaller than that size. This is analogous to a law that treats businesses smaller than a certain size differently than larger ones. The parameter L could be set initially to be equal to roughly 100,000 AGI tokens, for example.

The democratic mechanisms in the network are based on liquid (or delegative) democracy, meaning that when an agent A is qualified to vote on a decision, Agent A may also choose to delegate its voting power to some other Agent, Agent B. There may be smart contracts that allocate votes on some topics to some Agents based on metadata attached to the decisions or other more complex criteria. (For example, if you trust another Agent to do its due diligence on charitable projects, you may delegate to it your voting power for decisions about which projects are

beneficial, but for no other decisions.) The network will provide standard smart contracts to automatically delegate votes, but Agents can of course use any tools they wish for this purpose.

Major changes to the network will require more votes than minor changes. By major changes, we mean, for example,

- changes in the percentage of tokens allocated to different purposes (e.g., curation rewards versus benefit tokens),
- changes to how base reputation is calculated,
- changes to the quantitative parameters governing network economics,
- any decisions regarding creating more AGI tokens beyond those initially mined, or
- key design changes like moving to different blockchains and consensus algorithms.

By minor changes, we mean things like modifications to the APIs and ontologies used in inter-agent interactions.

For decisions regarding benefit tasks, the proposed mechanism will use a combination of votes by reputable Agents and benefit votes. The network gives benefit votes to Agents in proportion to their benefit quality ratings. (“Major changes” related to benefit tasks are changes to the system that certify tasks as benefit tasks.)

25.3.2 Transitioning to full democracy

In the early phases of network development, the Foundation will make some of the governance decisions. Decision-making will transition in phases to a purely democratic governance as the network matures, with the following specifics:

In years 1 and 2 of network operation (following the initial token issuance event), major changes are to be determined by the Foundation in accordance with the bylaws of the Foundation installed at the time of

network inception, while minor changes will be determined by a 51% majority of AGI token holders.

- In years 3 and 4
  - Major changes in the operation of SingularityNET: agreement of the Foundation plus a 51% majority of AGI token holder votes;
  - Minor changes in the operation of SingularityNET: a 51% majority of AGI token votes;
  - Major decisions related to benefit tasks: agreement of the Foundation plus 51% of AGI token votes plus 51% of benefit votes
- From year 5 onward
  - Major changes in the operation of SingularityNET: a 65% supermajority of AGI token votes;
  - Minor changes in the operation of SingularityNET: a 51% majority of AGI token votes;
  - Major decisions related to benefit tasks: a 65% supermajority of AGI token votes plus 65% of benefit votes are required.

26

TENCENT
ARCC: AI ETHICAL FRAMEWORK

26.1 Introduction

"Tencent is a world-leading internet and technology company that develops innovative products and services to improve the quality of life of people around the world." [476] Tencent is one of the "big four" of the Chinese media giants (BATH: Baidu, Alibaba, Tencent, Huawei). Tencent as most large international companies in China reports on ESG/CSR criteria[477], which is relevant for business ethics.

The Editors

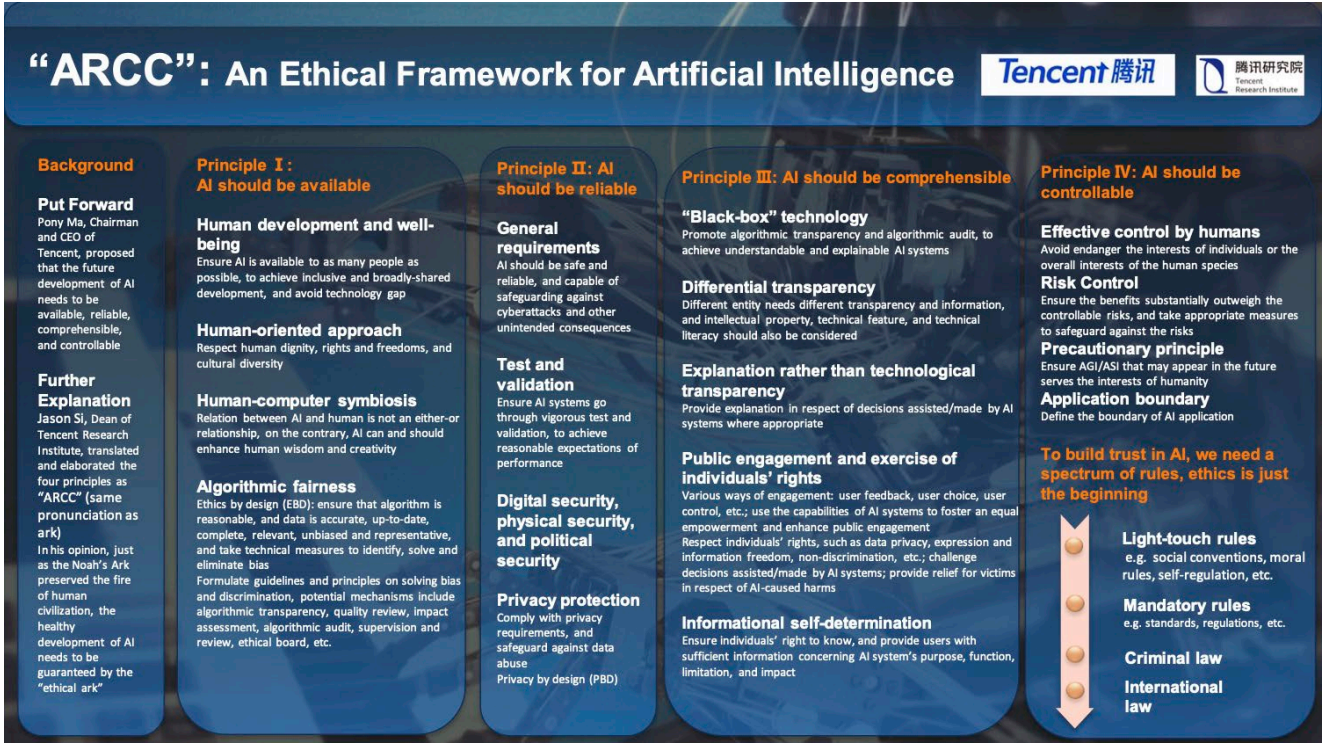
26.2 AI Available, reliable, comprehensive, controllable

On the specific AI ethics, Tencent developed the AI ARCC framework: AI should be available, reliable, comprehensive, controllable. [478]

[476] https://www.tencent.com/en-us/

[477] https://www.tencent.com/en-us/esg.html#respon-con-6

[478] https://www.tisi.org/13747.


"ARCC": An Ethical Framework for Artificial Intelligence
Tencent 腾讯
腾讯研究院 Tencent Research Institute
Background
Put Forward
Pony Ma, Chairman and CEO of Tencent, proposed that the future development of AI needs to be available, reliable, comprehensible, and controllable
Further Explanation
Jason Si, Dean of Tencent Research Institute, translated and elaborated the four principles as "ARCC" (same pronunciation as ark)
In his opinion, just as the Noah's Ark preserved the fire of human civilization, the healthy development of AI needs to be guaranteed by the "ethical ark"
Principle Ⅰ: AI should be available
Human development and well-being
Ensure AI is available to as many people as possible, to achieve inclusive and broadly-shared development, and avoid technology gap
Human-oriented approach
Respect human dignity, rights and freedoms, and cultural diversity
Human-computer symbiosis
Relation between AI and human is not an either-or relationship, on the contrary, AI can and should enhance human wisdom and creativity
Algorithmic fairness
Ethics by design (EBD): ensure that algorithm is reasonable, and data is accurate, up-to-date, complete, relevant, unbiased and representative, and take technical measures to identify, solve and eliminate bias
Formulate guidelines and principles on solving bias and discrimination, potential mechanisms include algorithmic transparency, quality review, impact assessment, algorithmic audit, supervision and review, ethical board, etc.
Principle Ⅱ: AI should be reliable
General requirements
AI should be safe and reliable, and capable of safeguarding against cyberattacks and other unintended consequences
Test and validation
Ensure AI systems go through vigorous test and validation, to achieve reasonable expectations of performance
Digital security, physical security, and political security
Privacy protection
Comply with privacy requirements, and safeguard against data abuse
Privacy by design (PBD)
Principle Ⅲ: AI should be comprehensible
"Black-box" technology
Promote algorithmic transparency and algorithmic audit, to achieve understandable and explainable AI systems
Differential transparency
Different entity needs different transparency and information, and intellectual property, technical feature, and technical literacy should also be considered
Explanation rather than technological transparency
Provide explanation in respect of decisions assisted/made by AI systems where appropriate
Public engagement and exercise of individuals' rights
Various ways of engagement: user feedback, user choice, user control, etc.; use the capabilities of AI systems to foster an equal empowerment and enhance public engagement
Respect individuals' rights, such as data privacy, expression and information freedom, non-discrimination, etc.; challenge decisions assisted/made by AI systems; provide relief for victims in respect of AI-caused harms
Informational self-determination
Ensure individuals' right to know, and provide users with sufficient information concerning AI system's purpose, function, limitation, and impact
Principle Ⅳ: AI should be controllable
Effective control by humans
Avoid endanger the interests of individuals or the overall interests of the human species
Risk Control
Ensure the benefits substantially outweigh the controllable risks, and take appropriate measures to safeguard against the risks
Precautionary principle
Ensure AGI/ASI that may appear in the future serves the interests of humanity
Application boundary
Define the boundary of AI application
To build trust in AI, we need a spectrum of rules, ethics is just the beginning
Light-touch rules
e.g. social conventions, moral rules, self-regulation, etc.
Mandatory rules
e.g. standards, regulations, etc.
Criminal law
International law

Background

Pony Ma, Chairman and CEO of Tencent, proposed that the future development of AI needs to be available, reliable, comprehensible, and controllable.

Further explanation

Jason Si, Dean of Tencent Research Institute, translated and elaborated the four principles as "ARCC" (same pronunciation as ark). In his opinion, just as the Noah's Ark preserved the fire of human civilization, the healthy development of AI needs to be guaranteed by the "ethical ark".

Principle I: AI should be available

Human development and well-being

Ensure AI is available to as many people as possible, to achieve inclusive and broadly-shared development, and avoid technology gap.

Human-oriented approach

Respect human dignity, rights and freedoms, and cultural diversity.

Human-computer symbiosis

Relation between AI and human is not an either-or relationship, on the contrary, AI can and should enhance human wisdom and creativity.

Algorithmic fairness

Ethics by design (EBD): ensure that algorithm is reasonable, and data is accurate, up-to-date, complete, relevant, unbiased and representative, and take technical measures to identify, solve and eliminate bias.

Formulate guidelines and principles on solving bias and discrimination, potential mechanisms include algorithmic transparency, quality review, impact assessment, algorithmic audit, supervision and review, ethical board, etc.

Principle II: AI should be reliable

General requirements

AI should be safe and reliable, and capable of safeguarding against cyberattacks and other unintended consequences.

Test and validation

Ensure AI systems go through vigorous test and validation, to achieve reasonable expectations of performance.

Digital security, physical security, and political security

Privacy protection

Comply with privacy requirements, and safeguard against data abuse Privacy by design (PBD).

Principle III: AI should be comprehensible

"Black-box" technology

Promote algorithmic transparency and algorithmic audit, to achieve understandable and explainable AI systems.

Differential transparency

Different entity needs different transparency and information, and intellectual property, technical feature, and technical literacy should also be considered.

Explanation rather than technological transparency

Provide explanation in respect of decisions assisted/made by AI systems where appropriate.

Public engagement and exercise of individuals' rights

Various ways of engagement: user feedback, user choice, user control, etc.; use the capabilities of AI systems to foster an equal empowerment and enhance public engagement.

Respect individuals' rights, such as data privacy, expression and information freedom, non-discrimination, etc.; challenge decisions assisted/made by AI systems; provide relief for victims in respect of AI-caused harms.

Informational self-determination

Ensure individuals' right to know, and provide users with sufficient information concerning AI system's purpose, function, limitation, and impact.

Principle IV: AI should be controllable. Effective control by humans

Effective control by humans

Avoid endanger the interests of individuals or the overall interests of the human species.

Risk Control

Ensure the benefits substantially outweigh the controllable risks, and take appropriate measures to safeguard against the risks.

Precautionary principle

Ensure AGI/ASI that may appear in the future serves the interests of humanity.

Application boundary

Define the boundary of AI application.

To build trust in AI, we need a spectrum of rules, ethics is just the beginning

Light-touch rules e.g. social conventions, moral rules, self-regulation, etc.

Mandatory rules e.g. standards, regulations, etc.

Criminal law.

International law.

27

HUAWEI
AI ETHICS AND THE FUTURE

27.1 Introduction

Huawei[479] is one of the "big four" of the Chinese media giants (BATH: Baidu, Alibaba, Tencent, Huawei). Huawei offers manifold products and services on all continents, from internet infrastructure to integrated e-campus of universities, from e-hospital equipment to future self-driving cars. AI is already a component in most of these business activities. The following text (27.2) reflects the ethical values of Huawei-AI. The next text (27.3) [480] is an excerpt on added value by AI Innovation. Both authors are in leading positions in Huawei on global level.

The Editors

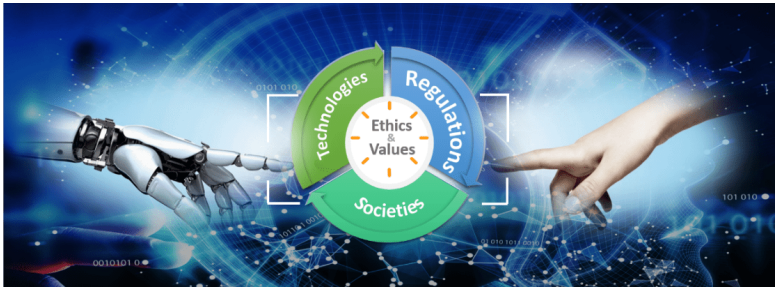
---

[479] Medhat Mahmpoud, *The Ethics and Values at the Core of AI*, Blog on Huawei Website, 25 May 2023. https://blog.huawei.com/2023/05/25/ethics-values-ai/. The author is is a core member in Huawei's global Digital Transformation Office and serves for Huawei in USA, Africa and Middle East.

[480] Zhou Hong, *Hypothesis and Visions for an Intelligent World.* https://www.huawei.com/en/huaweitech/publication/202301/hypotheses-visions-intelligent-world. The author is President of the Institute of Strategic Research of Huawei.

27.2 The ethics and values at the core of AI

After more than 50 years of intensive Artificial intelligence (AI) development, it has rapidly become one of the fastest growing technologies globally. AI holds the promise of huge transformative power that could produce significant socioeconomic and environmental benefits, productivity gains, and growth and prosperity for society – if it is developed wisely.



27.2.1 AI: The Good, the Bad, and the Ugly

By definition, AI involves programming high-performance computers to think, learn, and behave in a similar way to humans through a complex system of algorithms. AI is designed to help people make better predictions and informed decisions, and perform routine and complex tasks better than humans. Potentially, this could have huge benefits for a wide-range of sectors, including healthcare, mobility, agriculture, education, and communication.

- **Mobility:** Intelligent transportation will make traffic safer, faster, more efficient, and greener.
- **Healthcare:** Assisted diagnoses, early prevention, and precision treatment will have huge benefits for people's health.

- **Education:** AI opens the door to personalized learning; bridging the skills gaps; and optimizing workloads for teachers, student and administrators.
- **Agriculture:** Potential benefits include increased agricultural efficiency, improved crop yields, traceability, and lower food production costs
- **Manufacturing**: Improved productivity and lower costs will span every step of the manufacturing process.
- **Cross-culture communications**: Real-time translation across multiple languages will make communication easier than ever before.

However, alongside its potential, AI is fuelling anxieties, including legal and ethical concerns about the trustworthiness of AI systems. The possibilities of the risks it poses on society include cultural values, privacy, fraud, unemployment, and bias.

If left to its own devices free from human judgment or intervention, AI-generated errors can lead to devastating consequences, especially in critical and sensitive fields such as healthcare, traffic safety, investment decisions, cyberbullying, economic shifts, and job losses.

27.2.2 Who's in charge?

No single actor can be accountable for or provide answers to all the concerns about how to address the risks and challenges posed by AI, especially given its meteoric evolution. As a shared responsibility among all stakeholders, the good news is that international and national initiatives are already underway to establish the necessary technological and strategic frameworks and guidelines to ensure that AI is an ethically, socially, legally and ecologically responsible technology.

The objectives are designed to govern AI evolution with key principles to ensure the transparency, accountability, lawfulness, privacy, safety, human dignity, well-being, and environmental sustainability.

They also seek to ensure that AI prioritizes the public good (AI for Good) over commercial and geopolitical gains.

Because of the huge complexity of AI systems, those objectives can't be met unless close collaboration is established among three key actors:

1) Technology and Industry

2) Regulators

3) Academia and Education Institutions

1. Technology and industry: AI development platforms and their underlying technologies should offer a transparent, robust, trustworthy, secure, and safe AI infrastructure with the necessary mechanisms and tools to address data privacy, right to access, accountability and data protection. AI development infrastructure should enable developers and data scientists of different skill levels to rapidly build, train, test and deploy models, including data preparation, model development, optimization, and deployment. Other responsibilities include creating trustworthy algorithms with all the required tools, using secure and robust computing power, cloud infrastructure, and high-speed networking with sufficient storage capabilities.

2. Regulations: Global initiatives are in progress to generate corresponding AI frameworks, strategies, and guidelines to ensure responsible AI objectives will be met throughout the AI lifecycle with key three requirements to be incorporated: iii) Lawfulness: complying with applicable laws and regulations; ii) Ethical: ensuring adherence to principles and values; and iii) Robustness: both from a technical and social perspective.

Ideally, these three requirements must be coordinated in harmony with each other and key stakeholders to generate a pathway for responsible AI without over-regulation, as the technology is still evolving and needs to be embraced and adjusted over the years to come.

3. Education and academia: Academia and research will play an essential role for steering AI in the right direction for the common good.

Education institutions will be at the centre of developing AI's educational path, delivering the next generation of developers and researchers to place more emphasis on being ethically and value-driven.

Creating an educational atmosphere with the latest AI technology infrastructure and innovation hubs available can enable researchers and developers to work more closely with industry and communities through integrated knowledge-based AI-ecosystem platforms. This will help ensure that the design and development of AI systems is meaningful, adaptable, responsible, and robust enough to be trusted by the outside world.

It will also help universities and education institutions to generate adaptable policies and guidelines for using AI and continually re-validate these policies as technologies evolve.

New AI technologies can be quickly evaluated and used to facilitate learning, promote creativity, and thinking to find ways to embrace or integrate AI into teaching methodologies, while also maintaining academic standards and integrity. Teachers might redesign their courses, implementing more oral exams, promoting creativity through team work and handwritten essays, or integrating AI into classes by asking students to challenge its outputs.

27.2.3 A call for more collaboration and global dialogue

As AI has emerged as a top priority for all nations, international collaboration in generating guidelines and regulations are, now more than ever, becoming necessary to address the challenges presented by AI in terms of ethics, lawfulness, trustworthiness, and broader philosophical issues.

Technology, industry, academia, and regulators need to work together to establish best practices for universally accepted standards to maximize the potential of AI potential and mitigate its potential risks. Ensuring the harmony of technologies, industry, society, and regulations can in turn ensure that AI is beneficial to all.

It is certain that **education will definitely play a key role** by educating future AI professionals, students, younger generations, and society on the importance of accountability and the ethical considerations in AI development and applications.
Ethics and values have to be embedded in AI development by design.

Keeping humanity in the AI-loop must be thoroughly considered. Concepts such as swarm AI, for example, can amplify intelligence by building connected intelligent systems comprising people. Converging the power of algorithms with human expertise, knowledge, creativity, and intuition can enable people to bring unique ideas, wisdom, and insights to benefit AI's outcomes.

AI systems need to be **human-centric** based on a commitment to use AI for the service of humanity and the common good, supporting global cooperation, and the achievement of the SDGs with the objective of improving humanity's quality of life, while leveraging human knowledge, wisdom, values, morals, and sensibilities.

27.3 Visions and more values by applications

27.3.1 Three challenges AI faces in understanding the world

AI capabilities are improving rapidly, and so we need to consider how to ensure AI development progresses in a way that benefits all people and ensures that AI execution is accurate and efficient. In addition to ethics and governance, AI also faces three big challenges from a theoretical and technical perspective: AI goal definition, accuracy and adaptability, and efficiency.

The first challenge AI faces is that there is no agreed upon definition of its goals. What kind of intelligence do we need?

If there is no clear definition, it is difficult to ensure that the goals of AI and humanity will be aligned and to make reasonable measurements and classifications and scientific computations. Professor Adrian Bejan,

a physicist at Duke University, summarizes more than 20 goals for intelligence in his book The Physics of Life, including understanding and cognitive ability, learning and adaptability, and abstract thinking and problem-solving ability. There are many schools of AI, which are poorly integrated. One important reason for this is there are no commonly agreed upon goals for AI.

The second challenge AI faces is accuracy and adaptability. Learning based on statistical rules extracted from big data often results in non-transparent processes, unstable results, and bias. For example, when recognizing a banana using statistical and correlation-based algorithms, an AI system can be easily affected by background combinations and tiny noises. If other pictures are put next to it, the banana may be recognized as an oven or a slug. These pictures can be easily recognized by people, but AI makes these mistakes and it is difficult to explain or debug them.

The third challenge for AI is efficiency. According to the 60th TOP500 published in 2022, the fastest supercomputer is Frontier, which can achieve 1,102 PFLOPS while using 21 million watts of energy. Human brains, in contrast, can deliver about 30 PFLOPS with just 20 watts. These numbers show that the human brain is about 30,000 to 100,000 times more energy efficient than a supercomputer.

In addition to energy efficiency, data efficiency is also a major challenge for AI. It is true that we can better understand the world by extracting statistical laws from big data. But can we find logic and generate concepts from small data, and abstract them into principles and rules?

27.3.2 Breakthroughs and visions

We have come up with several hypotheses to address these three challenges:

- Knowledge comprises the concepts, attributes, relationships, and rules induced and abstracted from the facts and phenomena that we find in the external environment and ourselves.

- Intelligence is the ability to achieve goals in complex environments with limited resources through sensing and interaction, computing, and trial-and-error.
- We can achieve intelligence by extracting probability distributions from existing big data for model fitting and derivation, as well as by using deduction, hypotheses, and trial-and-error to ask questions and find creative solutions with little data or even no data at all.

Starting from these hypotheses, we can begin to take more practical steps to develop knowledge and intelligence.

At Huawei, our first vision is to combine systems engineering with AI to develop accurate, autonomous, and intelligent systems. In recent years, there has been a lot of research in academia about new AI architectures that go beyond transformers.

We can build upon these thoughts by focusing on three parts: perception and modelling, automatic knowledge generation, and solutions and actions. From there, we can develop more accurate, autonomous, and intelligent systems through multimodal perception fusion and modelling, as well as knowledge and data-driven decision-making.

Perception and modelling are about representations and abstractions of the external environment and ourselves. Automatic knowledge generation means systems will need to integrate the existing experience of humans into strategy models and evaluation functions to increase accuracy. Solutions can be directly deduced based on existing knowledge as well as internal and external information, or through trial-and-error and induction. We hope that these technologies will be incorporated into future autonomous systems, so that they can better support domains like autonomous driving networks, autonomous vehicles, and cloud services.

Our second vision is to create better computing models, architectures, and components to continuously improve the efficiency of intelligent computing. I once spoke with Fields Medallist Professor Laurent

Lafforgue about whether invariant object recognition could be made more accurate and efficient by using geometric manifolds for object representation and computing in addition to pixels, which are now commonly used in visual and spatial computing.

In their book Neuronal Dynamics, co-authors Gerstner, Kistler, Naud, and Paninski at École Polytechnique Fédérale de Lausanne (EPFL) explain the concept of functional columns in the cerebral cortex and the six-layer connections between these functional columns. It makes me wonder: Can such a shallow neural network be more efficient than a deep neural network?

A common bottleneck for today's AI computing is the memory wall. Reading, writing, and migrating data often takes 100-times more time than computing itself. So, can we possibly bypass conventional processors, instruction sets, buses, logic components, and memory components under von Neumann architecture, and redefine architectures and components based on advanced AI computing models instead?

27.3.3 Delivering more value with AI applications

Huawei has been exploring this idea by looking into the practical uses of AI. First, we have worked on "AI for Industry", which uses industry-specific large models to create more value. Industries face many challenges when it comes to AI application development. They need to invest a huge amount of manpower to label samples, find it difficult to maintain models, and lack the necessary capabilities in model generalization. Most simply they do not have the resources to do this.

To address these challenges, Huawei has developed L1 industry-specific large models based on its L0 large foundation models dedicated to computer vision, natural language processing, graph neural networks, and multi-modal interactions. These large models lower the barrier to AI development, improve model generalization, and address application fragmentation. The models are already being used to improve operational

efficiency and safety in major industries like electric power, coal mining, transportation, and manufacturing.

Huawei's Aviation & Rail Business Unit, for example, is working with customers and partners in Hohhot, Wuhan, Xi'an, Shenzhen, and Hongkong to explore the digital transformation of urban rail, railways, and airports. This has improved operational safety and efficiency, as well as user experience and satisfaction. The Shenzhen Airport has realized smart stand allocation with the support of cloud, big data, and AI, reducing airside transfer bus passenger flow by 2.6 million every year. The airport has become a global benchmark in digital transformation.

"AI for Science" is another initiative that will be able to greatly empower scientific computing. One example of this in action is the Pangu meteorology model we developed using a new 3D transformer-based coding architecture for geographic information and a hierarchical time-domain aggregation method. With a prior knowledge of global meteorological phenomena, the Pangu model uses more accurate and efficient learning and reasoning to replace time series solutions of hyperscale partial differential equations using traditional scientific computing methods. The Pangu model can produce 1-hour to 7-day weather forecasts in just a few seconds, and its results are 20% more accurate than forecasts from the European Centre for Medium-Range Weather Forecasts.

AI can also support software programming. In addition to using AI to do traditional retrieval and recommendation in a large amount of existing code, Huawei is developing new model-driven and formal methods. This is especially important for large-scale parallel processing, where many tasks are intertwined and correlated. Huawei has developed a new approach called Vsync which realizes automatic verification and concurrent code optimization of operating system kernels, and improves performance without undermining reliability. The Linux Community once discovered a difficult memory barrier bug which took community experts

more than two years to fix. With Huawei's Vsync method, however, it would have taken just 20 minutes to discover and fix the bug.

We have also been studying new computing models for automated theorem proving. Topos theory, for example, can be used to research category proving, congruence reasoning systems, and automated theorem derivation to improve the automation level of theorem provers. In doing this, we want to solve state explosion and automatic model abstraction problems and improve formal verification capabilities.

Finally, we are also exploring advanced computing components. We can use the remainder theorem to address conversion efficiency and overflow problems in real-world applications. We hope to implement basic addition and multiplication functions in chips and software to improve the efficiency of intelligent computing.

As we move towards the intelligent world, networks and computing are two key cornerstones that underpin our shift from narrow AI towards general-purpose AI and super AI. To get there, we will need to take three key steps. First, we will need to develop AI theories and technologies, as well as related ethics and governance, so that we can deliver ubiquitous intelligent connectivity and drive social progress. Second, we will need to continue pushing our cognitive limits to improve our ability to understand and control intelligence. Finally, we need to define the right goals and use the right approaches to guide AI development in a way that truly helps overcome human limitations, improve lives, create matter, control energy, and transcend time and space. This is how we will succeed in our adventure into the future.

****

Globethics

Globethics is an ethics network of teachers and institutions based in Geneva, with an international Board of Foundation and with ECOSOC status with the United Nations. Our vision is to embed ethics in higher education. We strive for a world in which people, and especially leaders, are educated in, informed by and act according to ethical values and thus contribute to building sustainable, just and peaceful societies.

The founding conviction of Globethics is that having equal access to knowledge resources in the field of applied ethics enables individuals and institutions from developing and transition economies to become more visible and audible in the global discourse.

In order to ensure access to knowledge resources in applied ethics, Globethics.net has developed four resources:

Globethics Library
The leading global digital library on ethics with over 8 million documents and specially curated content

Globethics Publications
A publishing house open to all the authors interested in applied ethics and with over 190 publications in 15 series

Globethics Academy
Online and offline courses and training for all on ethics both as a subject and within specific sectors

Globethics Network
A global network of experts and institutions including a Pool of experts and a Consortium

Globethics.net provides an electronic platform for dialogue, reflection and action. Its central instrument is the website:

https://globethics.net ■

Globethics Publications

The list below is only a selection of our publications. To view the full collection, please visit our website.

All products are provided free of charge and can be downloaded in PDF form from the Globethics library and at https://globethics.net/publications. Bulk print copies can be ordered from *publications@globethics.net* at special rates for those from the Global South.

Prof. Dr Fadi Daou, Executive Director. Prof. Dr Amélé Adamavi-Aho Ékué, Academic Dean, Dr Ignace Haaz, Managing Editor. M. Jakob Bühlmann Quero, Editor Assistant.

Find all Series Editors: https://globethics.net/publish-with-us
Contact for manuscripts and suggestions: *publications@globethics.net*

Sustainability Series

Philipp Öhlmann & Julianne Stork (Eds.), *Religious Communities and Ecological Sustainability in Southern Africa and Beyond*, 2024, 340pp. ISBN: 978-2-88931-548-2

Theses Series

Michael Heumann, *Zeitgenössische Wachstumskritik aus wirtschafts-philosophischer Perspektive: Zur Rekonstruktion der Wachstumsdynamik als Verselbständigung cartesischen Denkens*, 2023, 583pp. ISBN: 978-2-88931-542-0

Yosra Ben Ameur, *Essai sur la relation entre l'éthique et le droit des affaires. Partie I : La réception de l'éthique par le droit des affaires*, 2024, 438pp. ISBN : 978-2-88931-561-1

Yosra Ben Ameur, *Essai sur la relation entre l'éthique et le droit des affaires. Partie II : La protection de l'éthique par le droit des affaires,* 2024, 655pp. ISBN: 978-2-88931-565-9

Higher Education Series

Education Ethics

Divya Singh / Christoph Stückelberger (Eds.), *Ethics in Higher Education Values-driven Leaders for the Future*, 2017, 367pp. ISBN: 978–2–88931–165–1

Obiora Ike / Chidiebere Onyia (Eds.) *Ethics in Higher Education, Foundation for Sustainable Development*, 2018, 645pp. IBSN: 978-2-88931-217-7

Erin Green / Divya Singh / Roland Chia (Eds.), *AI and Ethics and Higher Education Good Practice and Guidance for Educators, Learners, and Institutions*, 2022, 324pp. ISBN 978-2-88931-442-3

Amélé Adamavi-Aho Ekué, Divya Singh, and Jane Usher (Eds.), *Leading Ethical Leaders: Higher Education Institutions, Business Schools and the Sustainable Development Goals*, 2023, 626pp. ISBN 978-2-88931-521-5

Research Ethics

Michelle Bergadàa, *Academic Plagiarism: Understanding It to Take Responsible Action*, 2021, 281pp. ISBN: 978-2-88931-330-3

Michelle Bergadàa, Paulo Peixoto (Eds.), *Academic Integrity: A Call to Research and Action*, 2023, 641pp. ISBN: 978-2-88931-517-8

Michelle Bergadaà / Paulo Peixoto (Eds.), *The New Boundaries of Academic Integrity*, 2024, 233p. ISBN 978-2-88931-580-2

Journal of Ethics in Higher Education

The focus and scope of JEHE is to answer to the request made by many faculty members from Globethics Consortium of higher education institutions, Network, Partners, Regional Programmes and participants to Globethics International Conferences to have a new space on Globethics platform for the publication of their research results in a scientific Journal.

https://jehe.globethics.net

ISSN: 2813-4389

This is only a selection of our latest publications, to view our full collection please visit:

https://globethics.net/publications

ISBN 978-2-88931-609-0

AI Governance Ethics

Artificial Intelligence with Shared Values and Rules

Artificial Intelligence influences already almost all sectors of society, from production, consumption and recycling of products to education and media, from agriculture to military. The search for AI governance from voluntary standards to international convention and standards is fast and urgent.

This book unites contributions of authors from all continents. It also offers a representative collection of official recommendations and guidelines on AI governance from national governments, continental entities like EU and AU to UN-organisations.

Editors

Christoph Stückelberger, Professor of Global Ethics, Founder and Honorary President of Globethics, Geneva/Switzerland

Maria Merchàn Rocamora, Jurist and Economist, Policy Support Officer Standards and Market Intelligence at Interpeace, former consultant at UN Alliance of Civilizations UNAOG

Divya Singh, Associate Professor of Law, lawyer, Advocate of the Higher Court of South Africa, Chief Academic Officer at STADIO Holdings, Cape Town, South Africa.

Pavan Duggal, Supreme Court Advocate, New Delhi/India, international Cyber Law Expert, Vice-Chancellor of Cyberlaw University, Delhi.

Globethics