# Human Rights Alignment: The Challenge Ahead for AI Lawmakers


**Marc Rotenberg**

**Abstract** The frameworks for the governance of AI have evolved rapidly. From the 2018 Universal Guidelines for AI on through the 2019 OECD/G20 AI Principles 2019, and the 2021 UNESCO Recommendation on AI Ethics, governments have agreed to the basic norms to regulate AI services. Two important legal frameworks are also now underway—the EU AI Act and the Council of Europe AI Convention. As these frameworks have evolved, we see the scope of AI governance models expand. From an initial focus on "human-centric and trustworthy AI" through the recognition of "fairness, accuracy, and transparency" as building blocks for AI governance, we see now consideration of sustainability, gender equality, and employment as key categories for AI policy. AI laws also overlap with familiar legal topics such as consumer protection, copyright, national security, and privacy. Throughout this evolution, we should consider whether the evolving models for the governance of AI are aligned with the legal norms that undergird democratic societies—fundamental rights, democratic institutions, and the rule of law. For democracies to flourish in the age of artificial intelligence, this is the ultimate alignment challenge for AI.

## 1 Introduction

As a field of study, digital humanism asks us to consider how we understand the impact of digital technologies on society and the humans who comprise it. In this chapter, we examine how societies respond to these challenges through legal and political institutions. The responses of national governments and international organizations to the specific challenges of the governance of AI become a key test of our ability to manage new technologies for social benefit. All of these undertakings begin with the premise that there must be a system of laws to safeguard fundamental rights, and in this respect, they go beyond the calls for ethical AI and responsible

M. Rotenberg (✉)
Center for AI and Digital Policy, Washington, DC, USA
e-mail: marc@caidp.org



H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_38

AI. At the same time, laws are imperfect. Language is imprecise. Technologies evolve rapidly. Powerful companies will resist constraints. And there is a risk that a dialectic process between technology and law could lead to outcomes that fail to protect well-established fundamental rights. This could occur, for example, if proponents of new technologies claim that well-established human rights, such as the protection of human dignity, autonomy, and privacy, must necessarily be altered to allow for the development of technology. For this reason, the articulation of norms for the governance of AI provides insight also to the ability of society to control the development of technology and to ensure that digital technology remains human-centric.

## 2 Main Part: Basic Concepts/Definitions/Methods and Critical Reflection

This section explores the development of legal frameworks for the governance of artificial intelligence. Law is generally understood to mean a system of rules that govern conduct. In democratic societies, law is derived from public debate and discussion and aims to reflect the will of the people, recognizing also the need to safeguard fundamental rights through constitutional limits on majority will.

Governance frameworks may also include influential policy frameworks (such as the Universal Guidelines for AI described below) as well as global agreements, which would include the OECD AI Principles,[1] the G20 AI Guidelines,[2] and the UNESCO Recommendation on AI Ethics.[3] These frameworks provide the basis for legal standards and international agreements and shape the conduct of those who develop and deploy AI systems as well as those who are subject to the outputs of AI systems.

Across the various governance frameworks, several key terms reoccur. These include "fairness," "accuracy," and "transparency," as well as "human-centric" and "trustworthy." These terms might also be considered the building blocks of AI law as they set out foundational values on which more specific direction is provided.

There are now several frameworks for the governance of artificial intelligence.

[1] OECD AI Principles (2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[2] G20 AI Guidelines (2019), https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf

[3] UNESCO **Recommendation on the Ethics of Artificial Intelligence** (2019)**, https://unesdoc.unesco.org/ark:/48223/pf0000380455**

## 2.1 Universal Guidelines for AI

*The Universal Guidelines for Artificial Intelligence* (UGAI)[4] were announced at the 2018 International Data Protection and Privacy Commissioners Conference at Brussels, Belgium—one of the most significant meetings of technology leaders and data protection experts in history. "The Guidelines are intended to maximize the benefits of AI, to minimize the risk, and to ensure the protection of human rights."[5] The UGAI incorporates elements of human rights doctrine, data protection law, and ethical guidelines. The Guidelines include several well-established principles for AI governance and put forward new principles not previously found in similar policy frameworks.[6] The Explanatory Memorandum states that the guidelines are primarily concerned with those systems that impact the rights of people.

According to the UGAI, the term artificial intelligence is both broad and imprecise and encompasses a variety of technological aspects which requires some degree of automated decision-making. The UGAI uses the term "guidelines" as a means of providing directional practices that can be useful for both governments and the private sector and recommends that the application of the guidelines should be incorporated into "ethical standards, adopted in national law and international agreements, and built into the design of systems."[7]

The UGAI is structured on 12 fundamental principles of right to transparency; right to human determination; identification obligation; fairness obligation; assessment and accountability obligation; accuracy, reliability, and validity obligation; data quality obligation; public safety obligation; cybersecurity obligation; prohibition on secret profiling; prohibition on unitary scoring; and a termination obligation. The UGAI also sets out prohibitions for mass surveillance and unitary (or social) scoring and includes a Termination obligation when it is no longer possible to maintain control of an AI system.

## 2.2 The OECD AI Principles/the G20 AI Guidelines (2019)

The OECD is an international organization that "works to build better policies for better lives." The goal of the OECD is to "shape policies that foster prosperity, equality, opportunity and well-being for all." The OECD emerged out of the Marshall Plan to assist Europe rebuild after the Second World War and to promote

[4] https://thepublicvoice.org/ai-universal-guidelines/

[5] https://thepublicvoice.org/ai-universal-guidelines/

[6] The Public Voice, *Explanatory Memorandum and References,* October 2018 https://thepublicvoice.org/ai-universal-guidelines/memo/

[7] *Id.*

economic interdependence. The OECD now has 38 member countries, spanning the Americas, Europe, and East Asia.[8]

The OECD led the global effort to develop and establish the most widely recognized framework for AI policy. This is a result of a concerted effort by the OECD and the member states to develop a coordinated international strategy. The OECD AI Principles also build on earlier OECD initiatives such as the OECD Privacy Guidelines, a widely recognized framework for transborder data flows and the first global framework for data protection.[9] The OECD AI Principles seek to promote AI that is innovative and trustworthy and respects human rights and democratic values.[10] The OECD set out five principles for the responsible stewardship of trustworthy AI:

1. Inclusive growth, sustainable development, and well-being
2. Human-centered values and fairness
3. Transparency and explainability
4. Robustness, security, and safety
5. Accountability

The OECD also set out five recommendations for national policies and international cooperation for trustworthy AI:

1. Investing in AI research and development
2. Fostering a digital ecosystem for AI
3. Shaping an enabling policy environment for AI
4. Building human capacity and preparing for labor market transformation
5. International cooperation for trustworthy AI

The OECD AI Principles were subsequently endorsed by the G20 nations in 2019. As a consequence, more than 50 countries have endorsed either the OECD AI Principles or the G20 AI Guidelines.

The remarks of the former OECD Secretary General Angel Gurria at the 2020 G-20 Digital Economy Ministers Meeting also provide insight into the work of the OECD on AI.[11] Secretary Gurria said, "AI's full potential is still to come. To achieve this potential, we must advance a human-centered and trustworthy AI, that respects the rule of law, human rights, democratic values and diversity, and that includes appropriate safeguards to ensure a fair and just society. This AI is consistent with the

[8] List of OECD Member countries - Ratification of the Convention on the OECD, https://www.oecd.org/about/document/ratification-oecd-convention.htm

[9] OECD, *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1981),* https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

[10] https://www.oecd.org/digital/artificial-intelligence/#

[11] CAIP Update 1.2, *OECD's Gurria Underscores AI Fairness at G-20* (July 26, 2020), https://dukakis.org/center-for-ai-and-digital-policy/center-for-ai-policy-update-oecds- gurria-underscores-ai-fairness-at-g-20-meeting/

G20 AI Principles you designed and endorsed last year, drawing from the OECD's AI Principles."

## 2.3 The UNESCO Recommendation on AI Ethics

In November 2021, the 193 member states of UNESCO adopted the Recommendation on the Ethics of Artificial Intelligence, the most comprehensive global framework to date for the governance of AI.[12] It will not only protect but also promote human rights and human dignity and will be an ethical guiding compass and a global normative bedrock allowing to build strong respect for the rule of law in the digital world.[13] UNESCO Director General Audrey Azoulay stated, "The world needs rules for artificial intelligence to benefit humanity. The recommendation on the ethics of AI is a major answer. It sets the first global normative framework while giving member states the responsibility to apply it at their level. UNESCO will support its 193 member states in its implementation and ask them to report regularly on their progress and practices."

The UNESCO Recommendation was the outcome of a multiyear process and was drafted with the assistance of more than 24 experts.[14] According to UNESCO, the "historical text defines the common values and principles which will guide the construction of the necessary legal infrastructure to ensure the healthy development of AI."[15] UNESCO explained, "The Recommendation aims to realize the advantages AI brings to society and reduce the risks it entails. It ensures that digital transformations promote human rights and contribute to the achievement of the Sustainable Development Goals, addressing issues around transparency, accountability and privacy, with action-oriented policy chapters on data governance, education, culture, labour, healthcare and the economy."

The UNESCO recommendation carried forward earlier principles for the governance of AI and also introduced new safeguards such as gender equity and sustainability. The key achievements of the UNESCO AI Recommendation include:

[12] UNESCO, *UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence* (Nov. 25, 2021), https://en.unesco.org/news/unesco- member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence

[13] UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (2021), https://unesdoc.unesco.org/ark:/48223/pf0000380455

[14] UNESCO, *Preparation of a draft text of the Recommendation: Ad Hoc Expert Group,* https://en.unesco.org/artificial-intelligence/ethics#aheg

[15] UNESCO, *UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence* (Nov. 25, 2021), https://en.unesco.org/news/unesco- member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence

1. **Protecting data.** The UNESCO Recommendation calls for action beyond what tech firms and governments are doing to guarantee individuals more protection by ensuring transparency, agency, and control over their personal data.
2. **Banning social scoring and mass surveillance**. The UNESCO Recommendation explicitly bans the use of AI systems for social scoring and mass surveillance.
3. **Monitoring and evaluation.** The UNESCO Recommendation establishes new tools that will assist in implementation, including ethical impact assessments and a readiness assessment methodology.
4. **Protecting the environment.** The UNESCO Recommendation emphasizes that AI actors should favor data, energy, and resource-efficient AI methods that will help ensure that AI becomes a more prominent tool in the fight against climate change and on tackling environmental issues.

The Recommendation aims to provide a basis to make AI systems work for the good of humanity, individuals, societies, and the environment and ecosystems and to prevent harm. It also aims at stimulating the peaceful use of AI systems. The Recommendation provides a universal framework of values and principles of the ethics of AI. It sets out four values: respect, protection, and promotion of human rights and fundamental freedoms and human dignity; environment and ecosystem flourishing; ensuring diversity and inclusiveness; and living in peaceful, just, and interconnected societies.

Further, the Recommendation outlines 10 principles—proportionality and do no harm, safety and security, fairness and nondiscrimination, sustainability, right to privacy and data protection, human oversight and determination, transparency and explainability, responsibility and accountability, awareness, and literacy—backed up by more concrete policy actions on how they can be achieved. The Recommendation also introduces red lines to unacceptable AI practices. For example, it states that "AI systems should not be used for social scoring or mass surveillance purposes."

The Recommendation focuses not only on values and principles but also on their practical realization, with concrete policy actions. The UNESCO Recommendation encourages member states to introduce frameworks for ethical impact assessments and oversight mechanisms. According to UNESCO, member states should ensure that harms caused through AI systems are investigated and redressed, by enacting strong enforcement mechanisms and remedial actions, to make certain that human rights and fundamental freedoms and the rule of law are respected.

## *2.4 The EU AI Act*

With the introduction of the Artificial Intelligence Act, the European Union aims to create a legal framework for AI to promote trust and excellence. The AI Act would establish a risk-based framework to regulate AI applications, products, and services. The rule of thumb: the higher the risk, the stricter the rule. The AI Act seeks to

protect fundamental rights and public safety. The legislation will also prohibit certain AI applications, such as social scoring and mass surveillance, as UNESCO has recently urged in the Recommendation on AI ethics, endorsed by 193 countries.[16]

In various comments to the European Parliament and the European Council, groups such as the Center for AI and Digital Policy have sought to align the EU AI Act with such frameworks as the Universal Guidelines for AI that underscore the need to protect fundamental rights. Some of the recommendations from CAIDP are as follows:

Prohibit Pseudoscientific and Discriminatory AI Systems

- *Require scientific validity for AI systems*
- *Ban predictive policing*
- *Ban emotion recognition systems*
- *Ban biometric categorization systems*
- *Apply bans to both public and private entities*

Safeguard Fundamental Rights

- *Remove the broad exclusions for law enforcement*
- *Remove the exclusions for ex ante systems*
- *Remove the national security exclusion*
- *Correct the unequal protection of asylum seekers and refugees*

Ensure Transparency and Accountability

- *Mandate ex ante impact assessments*
- *Record serious incidents*
- *Require private users to registers*
- *Mandate independent, third-party auditing*
- *Regulate general-purpose AI (GPAI) systems*
- *Establish obligation to terminate AI systems no longer under human control*

Protect Societal Interests

- *Protect the environment*
- *Safeguard disability rights*
- *Adopt UNESCO Recommendation on AI Ethics*

As the EU AI Act is still under development, it remains to be seen which of these recommendations will be adopted. The text adopted by the Parliament extended

[16] Center for AI and Digital Policy, EU Artificial Intelligence Act, https://www.caidp.org/resources/eu-ai-act/

prohibitions to subliminal techniques, biometric categorization, predictive policing, Internet-scrapped facial recognition databases, and emotion recognition.[17]

## 2.5 The Council of Europe AI Convention

The Council of Europe (COE) is the continent's leading human rights organization.[18] The COE is comprised of 47 member states, 27 of which are members of the European Union. All COE member states have endorsed the European Convention of Human Rights, a treaty designed to protect human rights, democracy, and the rule of law. Article 8 of the Convention, concerning the right to privacy, has influenced the development of privacy law around the world.

Several AI initiatives are underway at the Council of Europe, including at the Council of Ministers, the COE Parliamentary Assembly. Marija Pejčinović Burić, Secretary General of the Council of Europe, has said "It is clear that AI presents both benefits and risks. We need to ensure that AI promotes and protects our standards .... The Council of Europe has, on many occasions, demonstrated its ability to pioneer new standards, which have become global benchmarks."[19]

In October 2020, the Parliament Assembly of the Council of Europe adopted a new resolution on the Need for Democratic Governance of Artificial Intelligence.[20] The Assembly called for "strong and swift action" by the Council of Europe. The parliamentarians warned that "soft-law instruments and self-regulation have proven so far not sufficient in addressing these challenges and in protecting human rights, democracy and rule of law."

In a set of recommendations examining the opportunities and risks of AI for democracy, human rights, and the rule of law adopted in October 2020 as well, the Parliamentary Assembly called on the Committee of Ministers to take into account the particularly serious potential impact of the use of artificial intelligence "in policing and criminal justice systems"[21] or "on the enjoyment of the rights to

[17] Luca Bertuzzi, AI Act enters final phase of EU legislative process, Euractiv, June 14, 2023, https://www.euractiv.com/section/artificial-intelligence/news/ai-act-enters-final-phase-of-eu-legislative-process/

[18] Council of Europe, *Who we are,* https://www.coe.int/en/web/about-us/who-we-are

[19] Council of Europe, *Artificial intelligence and human rights,* https://www.coe.int/en/web/artificial-intelligence/secretary-general-marija-pejcinovic- buric

[20] Council of Europe, Parliamentary Assembly, *Need for democratic governance of artificial intelligence* (Oct. 22, 2020), https://pace.coe.int/en/files/28803/html

[21] Parliamentary Assembly, *Recommendation 2182(2020) Justice by algorithm – The role of artificial intelligence in policing and criminal justice systems* (Oct. 22, 2020) https://pace.coe.int/en/files/28806/html; See also, *Resolution 2342 (2020)* https://pace.coe.int/en/files/28805

equality and non-discrimination,"[22] when assessing the necessity and feasibility of an international legal framework for artificial intelligence.

At present, a draft text circulated by the Committee on AI of the Council of Europe seeks to establish a comprehensive global treaty for the governance of AI.[23] In a statement issued in May 2023, the Council of Europe Committee on Artificial Intelligence (CAI), explained, "The CAI is committed to ensuring that the Framework Convention will be human- centred, open to non-member States, and adopt a risk-based approach to the design, development, and use of AI systems facilitating the prevention of harmful uses of AI systems and promoting the use of this digital technology for the good of society, including by allowing for safe innovation."[24]

## *2.6 Challenges Ahead*

An ongoing challenge in AI policy concerns the ability to establish and enforce prohibitions on certain AI deployments. For example, the UNESCO Recommendation on AI Ethics discussed above establishes prohibitions on the use of AI techniques for social scoring and mass surveillance, yet many of the countries that have endorsed the UNESCO Recommendation continue to support the use and deployment of AI systems for these purposes. China, for example, continues to deploy a social credit system, based on AI, that is intended to align the private behavior of Chinese citizens with the political aims of the Chinese Communist Party.[25] Although there is some dispute as to the scope of the social credit system and a recognition that China needs to assess credit worthiness for efficient markets, the incorporation of certain factors in the evaluation, such as "picking quarrels and provoking trouble," is precisely the factors in an AI model that raise concern. The use of AI for remote biometric identification, a form of mass surveillance, remains a concern not only in China but in many countries that have installed camera systems for monitoring public spaces. Over time, these networks have become more sophisticated, providing the ability to link images to individuals and then to government profiles that may also provide risk assessments that lead to police intervention before any unlawful act has occurred. The effective governance of AI in democratic societies will require limitations and prohibitions on the deployment of such AI-driven systems

[22] Parliamentary Assembly, *Recommendation 2183 (2020) Preventing discrimination caused by the use of artificial intelligence* (Oct. 22, 2020) https://pace.coe.int/en/files/28809/html; See also, *Resolution 2343 (2020)* https://pace.coe.int/en/files/25318/html

[23] Center for AI and Digital Policy, Council of Europe AI Treaty, https://www.caidp.org/resources/coe-ai-treaty/

[24] Statement of the Council of Europe Committee on Artificial Intelligence (CAI), https://rm.coe.int/cai-statement-fr/1680ab6e85

[25] John Feng, How China's Social Credit System Works, Newsweek, Dec. 22, 2022, https://www.newsweek.com/china-social-credit-system-works-explained-1768726

There are also the legal frameworks currently underway at the European Union, the Council of Europe, and many governments around the world. There are challenges ahead for both adoption and effective implementation. There is currently a 2-year period from the time the EU AI Act is finalized to actual enforcement. Some are concerned that this gap will allow the use of unregulated AI systems that put at risk fundamental rights and public safety. However, a proposal from industry to develop an interim "AI pact" or "code of conduct" is opposed by civil society organizations as it would undermine democratic decision-making.[26] Regarding the Council of Europe Treaty, there are concerns also about implementation and the possibility that national governments will endorse the treaty nonetheless.

Finally, there remains an existential challenge—will humans remain in control of the AI systems they create? Stuart Russell has expressed this concern in *Human Compatible: Artificial Intelligence and the Problem of Control* in 2019. In recent years, there is growing attention to this issue, as new AI techniques challenge even the ability to deliberate. From this perspective, the ability to develop effective legislation to govern AI becomes even more critical.

## 3 Conclusions

- The governance structures for artificial intelligence have evolved rapidly, from framework principles to enforceable laws. Many of the governance structures emphasize "human-centric" and "trustworthy" AI.
- As the governance of AI has evolved, so too has the range of issues that fall within the AI domain. Early framework principles focused on automated decision-making and emphasized fairness, accountability, and transparency. More recent governance mechanisms set out principles for equity, public safety, and environmental sustainability.
- AI governance includes prohibitions on certain AI deployments such as social scoring, mass surveillance, and biometric categorization.
- Laws that govern AI interact with other legal rules, including consumer protection, copyright, data protection, national security, and privacy.
- One of the key challenges for AI governance concerns accountability: how to assess AI outcomes if it is not possible to determine how they are produced? A range of solutions has been proposed including explainability and traceability, certification, and transparency obligations.
- The biggest challenge for AI governance may simply be ensuring that AI is aligned with democratic values, fundamental rights, and the role of law.

[26] BEUC, **EU-US AI voluntary code of conduct and an 'AI Pact' for Europe, June 5, 2023, https://www.beuc.eu/letters/eu-us-ai-voluntary-code-conduct-and-ai-pact-europe**

**Discussion Questions for Students and Their Teachers**

1. What are some of the reasons to have laws to govern artificial intelligence?
2. What are the characteristics of AI governance frameworks?
3. Which AI "use cases" would you consider to be high risk and why? And which would you consider to be low risk and why?
4. Are there AI deployments that you would prohibit? If yes, why? If no, why not?
5. What recommendation would you make for an AI governance principle?

**Learning Resources for Students**

1. Daron Acemoglu and Simon Johnson, *Power and Progress: Our 1000-Year Struggle Over Technology and Prosperity* (Public Affairs Books 2023)

   As the authors explain, "A thousand years of history and contemporary history make one thing clear: progress depends on the choices we make about technology. New ways of organizing production can either serve the narrow interests of an elite or become the foundation for widespread prosperity." *Powers and Progress* include a detailed critique of the AI economy and recommendations for concrete actions.
2. Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology* (Oxford 2023)

   Bradford examines three competing approaches for the digital economy—the American market-driven model, the Chinese state-driven model, and the European rights-driven regulatory model. Which digital empire will prevail in the contest for global influence remains an open question, yet their contrasting strategies are increasingly clear and will have far-reaching consequences for the governance of artificial intelligence.
3. Center for AI and Digital Policy, *Artificial Intelligence and Democratic Values Index* (2023)

   A comprehensive review of AI policies and practices in 75 countries. It provides a methodology to compare country practices and assess trends across 12 key metrics and includes the text of the key AI policy frameworks, including the OECD AI Principles and the UNESCO Recommendation on AI Ethics.
4. European Law Institute, Guiding Principles for Automated Decision Making in the EU (2022)

   The Innovation Paper sets out 12 principles for automated decision-making. The ELI Guiding Principles include such novel concepts for AI governance such as traceability, reasoned decisions, risk allocation, and responsible ADM, including impact assessment on democratic values.
5. Marc Rotenberg, Time to Assess National AI Policies, Communications of the ACM, Nov. 24, 2020

   In this article for a computer science journal, the author explains the origins of the AI and Democratic Values report. "Our goal is to understand the commitments that governments have made, the AI initiatives they have launched, and the policies they have established to protect fundamental rights and to safeguard the public."

6. Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019)

   One of the world's leading AI researchers describes the challenges ahead to maintain control of artificial intelligence. Professor Russell proposes that we reassess the aims of AI systems, to build in uncertainty about pursuing outcomes and to ensure alignment with human preferences reflected in human behavior.