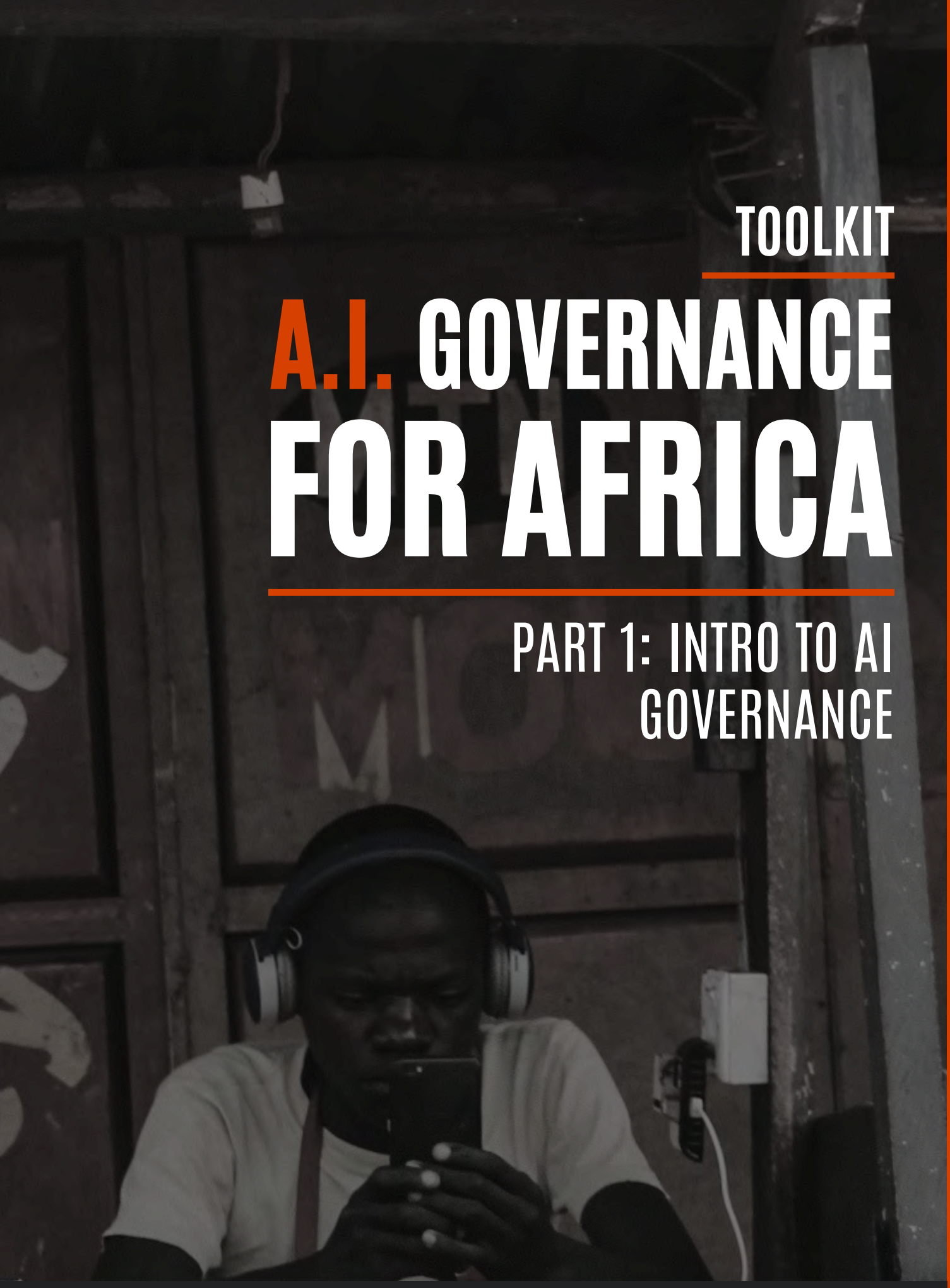

TOOLKIT

A.I. GOVERNANCE FOR AFRICA

PART 1: INTRO TO AI GOVERNANCE

Thomson Reuters Foundation

AI GOVERNANCE FOR AFRICA TOOLKIT SERIES

PART 1: INTRODUCTION TO AI GOVERNANCE FRAMEWORKS

December 2024

This is **Part 1** of the Thomson Reuters Foundation's toolkit series on AI Governance for Africa. It introduces AI governance principles and approaches, and outlines international frameworks, with case studies from the European Union, the United States, and China.

Part 2 will examine emerging AI governance instruments and approaches on the continent, with a focus on Southern Africa – in particular, South Africa, Zambia, and Zimbabwe.

Part 3 will explore options to build an advocacy strategy in pursuit of AI governance.

Published by the Thomson Reuters Foundation (trust.org)

Supported by the Patrick J. McGovern Foundation (mcgovern.org)

Revised in 2024 by S'lindile Khumalo and Murray Hunter from ALT Advisory (altadvisory.africa), with contributions in 2023 by Tara Davis and Tharin Pillay.

Contents

PART 1: INTRODUCTION TO AI GOVERNANCE FRAMEWORKS

Introduction

While there is no universal definition of Artificial intelligence (AI), UNESCO's Recommendation on the Ethics of AI describes AI systems as those "…**which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control**."[1]

These technologies have permeated our everyday. They are no longer confined to use by big tech or billionaires; ordinary individuals have AI in their pockets – and they are using it.

This growing prevalence of AI systems has accelerated the need for AI governance frameworks. The 2022 public release of ChatGPT – OpenAI's chatbot which is capable of generating novel content in response to prompts – is a prominent milestone in the public story of of AI,[2] with its release sparking a rapid growth in products and applications which make generative AI more accessible to ordinary people, and which have arguably made the uses and risks of AI technologies more visible in the public imagination.

Outside of large language models (LLMs) like ChatGPT, some everyday examples of the use of AI systems include rideshare apps, online banking systems, and e-commerce. Governments the world over are using AI for a wide range of activities including to build smart city platforms, bolster policing systems, and for service delivery, among others.

As AI systems become increasingly embedded in everyday life, they raise important questions about **regulation, ethics, and their potential impact on human rights**. These questions find form in debates about the governance of AI – how do we provide protection without stifling innovation? How can the law keep pace with the evolving nature of AI? Should AI be governed internationally or domestically?

Despite the complexity of AI governance, it is clearly a global concern. Countries are at different phases of resolving these questions and have implemented a range of governance instruments in response to concerns.

This toolkit series is a continuation and update of work the Thomson Reuters Foundation first undertook in 2023. It unpacks the context of AI governance, in Africa and globally, and considers advocacy approaches for future governance, with a particular focus on the Southern African context. It does so in the following ways:

- This part, **Part 1**, gives an introduction to AI governance principles and approaches, and outlines international frameworks, with case studies from the European Union (EU), the

[1] UNESCO "Recommendation on the Ethics of Artificial Intelligence" (2022). (Accessible here.)
[2] K Hu "ChatGPT sets record for fastest growing user base – analyst note" Reuters (1 February 2023). (Accessible here.)

United States, and China. In doing so it considers governance trends and important considerations included in governance instruments.

- **Part 2** examines existing and emerging AI governance instruments in Southern Africa – in particular, South Africa, Zambia, and Zimbabwe. More broadly, it also outlines continental responses and details existing governing measures in Africa.

- **Part 3** explores a series of key questions for the design of advocacy strategies on AI governance, particularly in African contexts.

Understanding governance

Governance instruments are the tools, mechanisms and strategies that are used to guide, regulate, and manage the various aspects of AI. Some of the instruments used include:

- **Guidelines and standards**: These are generally non-binding documents created by organisations, professional bodies and governments that provide a framework or set of principles to guide developers, users of AI or policy makers on important considerations such as fairness, transparency, accountability, and the avoidance of harm.

- **Government strategy:** A government strategy is a high-level plan or approach that outlines the goals, priorities, and actions that a government intends to take in regard to AI. Strategy documents don't have the binding force of laws but serve as a guide for understanding a government's intended response.

- **Policy:** A government policy is a more detailed and specific set of guidelines, rules, or principles that guide decision-making and actions in a particular area. Policies provide a roadmap for implementing a government's strategy. Depending on domestic law, some policies may be enforced through measures such as monitoring or audits and could incur penalty such as disciplinary action or the revocation of a benefit or license.

- **Law:** Laws or regulations are codified rules enacted by a legislative body. The rules are enforceable, and non-compliance can result in penalty. In the context of AI, a country could implement one single law to deal with all aspects of AI (as evidenced by the EU AI Act) or could take a fragmented approach where different laws regulate different aspects.

These instruments often interact with each other: government strategies guide the creation of policies, and policies can guide the drafting of laws to ensure that legal frameworks align with broader strategic goals. All of these instruments would likely consider the principles developed in guidelines and standards.

Why we need AI governance

Despite the great diversity of AI technologies, and their uses and contexts, there are a range of common concerns and risks associated with AI that underpin the need for governance. (The AI Risk Repository has identified and categorised more than 700 AI risks documented in academic research.)[3] This section explores some of the most prominent themes in AI risk.[4]

Discrimination and bias

There is significant concern of **algorithmic bias**, where AI systems reproduce biases in their design or in the data they are trained, leading to discriminatory or unfair outcomes (often based on race, gender, or other sensitive characteristics). For example:

- Certain AI recruitment tools used to screen job applicants have been found to favour men over women,[5] or younger applicants over older ones.[6]
- Banks in the United States which use a particular software to assess people applying for a home loan were found to be 80% more likely to reject black applicants than white applicants with a similar financial status.[7]

These biases are thought to be the result of the data used to train the AI systems. For example, if an AI recruitment software analysed all the hiring decisions previously made by human recruiters (who may have been more likely to give jobs to male candidates, or younger candidates), the AI system may replicate those patterns without meaning to. From a rights perspective, algorithmic bias undermines the right to equality and non-discrimination, which is protected under the International Bill of Human Rights.[8]

Lack of transparency or explainability

There are a broad range of concerns relating to **lack of transparency** in how AI systems are designed and trained, and the **lack of explainability** (sometimes called **interpretability**) in how these systems arrive at a particular decision or outcome. This compounds the discrimination risk described above. The decisions or output of an AI system can have far-reaching consequences (for example, a bank's assessment of a person's creditworthiness or a recruiter's assessment of

---

[3] P Slattery and others "A systematic evidence review and common frame of reference for the risks from artificial intelligence" AI Risk Repository (2024). (Accessible here.)
[4] This section draws on some of the themes outlined in the AI Risk Repository (above note 3).
[5] J Dastin "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters (October 2018). (Accessible here.)
[6] Equal Employment Opportunity Commission "EEOC Sues iTutorGroup for Age Discrimination" (May 2022). (Accessible here.)
[7] E Martinez and L Kirchner "The secret bias hidden in mortgage-approval algorithms" Associated Press (August 2021). (Accessible here.)
[8] The International Bill of Human Rights refers to the Universal Declaration of Human Rights (1948) (Accessible here), the International Covenant on Civil and Political Rights (1976) (Accessible here), and the International Covenant on Economic, Social and Cultural Rights (1966) (Accessible here.)

whether to give a person a job interview), but it may be impossible to know how this decision was made, and whether the decision was fair or correct.

Privacy and data protection concerns

There are several dimensions to **AI privacy concerns**, such as:

- 'Design-side' concerns, relating to how AI systems are 'trained' on vast amounts of information, which may include storage and analysis of people's personal information without their knowledge or informed consent.
- 'Use-side' concerns, relating to how AI systems could be used to harm people's privacy rights, for example by inadvertently leaking personal information which are stored by an AI system, or by AI-driven analysis of personal information, or by the exploitation of security vulnerabilities in an AI system.
- AI is at the heart of invasive surveillance technology, such as facial recognition or bulk analysis of communications data.

Since the right to privacy is internationally recognised, governance instruments should include principles and mechanisms protecting this right. This should include effective recourse channels for instances where private information is compromised.

Misinformation and disinformation

There are broad concerns at the **capacity for AI systems to produce or spread false or misleading information**, potentially on a grand scale, such as:

- The tendency for generative AI systems such as Chat GPT to confidently present incorrect information as fact (sometimes called 'hallucination').[9]
- The capacity for malicious actors to use AI to generate fake but realistic-looking images, video, or other media, to stoke false beliefs about the world – for example by depicting a politician doing something outrageous.[10]
- The risk of content-ranking and content-moderation algorithms on social media or search engine platforms enabling the spread of false or harmful content to a wider audience, or to artificially suppress valid information or opinions.[11]

Misinformation and disinformation may contribute to the violation of several rights including electoral rights, the right to equality and non-discrimination, and freedom of expression, among others.

[9] W Zhao and others "Wild Hallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries" [Preprint] (2024). (Accessible here.)
[10] M Adami "How AI-generated disinformation might impact this year's elections" Reuters Institute for the Study of Journalism (March 2024). (Accessible here.)
[11] See for example M Elswah "Does AI Understand Arabic? Evaluating The Politics Behind the Algorithmic Arabic Content Moderation" Carr Center for Human Rights Policy (2023). (Accessible here.)

Loss of human autonomy

These concerns broadly relate to broader social impacts from the rollout of AI systems, beyond the working of any specific technology, such as:

- Social and economic inequalities and **job losses** as human workers are replaced by increasingly sophisticated AI systems.
- **Over-dependence on AI decision-making**, where human operators accept the decisions or outputs of AI systems uncritically, or voluntarily give up decision-making powers to those systems because of a conscious or unconscious sense that the AI systems are infallible.

Environmental harms

There is a growing appreciation that advanced AI systems have a **significant carbon footprint**, as they require a huge amount of computing power and server infrastructure, as well as electricity usage (to power the infrastructure) and water (to prevent it from overheating). For example, between 2020 and 2023, the carbon emissions of major technology companies such as Microsoft, Google, and Meta, increased by 40 percent to 65 percent due to their investments in AI.[12] This has raised concerns that AI development is an obstacle to global responses to the climate crisis, and consequently, has a bearing on the enjoyment of the right to a healthy environment.

Geopolitical power imbalances

The fact that the majority of AI technology is developed and owned in the global North has raised concerns about **asymmetric development** in AI. The attendant risks include:

- Perpetuating global inequalities and excluding global South communities from potential social and economic benefits of AI technologies.
- Leaving global South communities, especially more marginalised groups, more vulnerable of the attendant risks of AI technologies, such as algorithmic bias and exclusion.

Existential AI risks

Perhaps most severely, there are genuine concerns that without sufficient guardrails, AI technologies could develop capabilities to bring about mass harm. For example:

- Certain AI technologies have already been found to be capable of using deception or manipulation to carry out the task they were designed for,[13] pointing to the need for strong ethical programming to ensure AI are guided by human values, laws, and goals.

---

[12] G Noble and F Berry "Power-hungry AI is driving a surge in tech giant carbon emissions" The Conversation (July 2024). (Accessible here.)
[13] P Park and others "AI deception: A survey of examples, risks, and potential solutions" (2024) 5 *Patterns*. (Accessible here.)

- Many AI developers and researchers predict that future AI technology may become self-aware, or start to pursue their own goals or interests in opposition to those of humankind.

The aim of AI governance efforts is to create harmonising frameworks that ensure AI technologies are developed in ways which address these risks while harnessing the potential benefits. Importantly, these governance efforts also aim to ensure similar standards apply to each developer and in each jurisdiction, so that everyone follows the same set of rules.

Principles of best practice

As demonstrated in this toolkit series, AI governance is scattered across various instruments at the international, regional, and even domestic levels. While there is no singular source for best practice principles, there is significant overlap across this range of documents, perhaps best summed up by the **UNESCO Recommendation on the Ethics of Artificial Intelligence**,[14] which outlines principles which should underscore any design, use, or output of AI:

- Proportionality and protection against harm
- Fairness and non-discrimination
- Safety and security
- Sustainability
- Privacy and data protection
- Human oversight and determination
- Transparency and explainability
- Responsibility and accountability

The social and political context

Countries across the world are scrambling to regulate AI. In understanding the regulatory approaches being adopted, we should examine the social and political context informing this. There are several factors currently influencing the creation of international AI governance.

- **Asymmetric AI development:** It costs hundreds of millions of dollars to train cutting-edge AI models, due to the immense amounts of computing power, data, and other raw materials required,[15] which limits who can create these models. At present AI development is overwhelming driven by private industry rather than academic or government institutions,[16] with the vast majority of advanced AI models coming from labs in the United States (like OpenAI, Google, Meta, and Anthropic); labs in China are in distant second place, with the

[14] Above n1.
[15] W Knight "OpenAI's CEO Says the Age of Giant AI Models Is Already Over" Wired (17 April 2023). (Accessible here.)
[16] Stanford University Institute for Human-Centered AI "The AI Index 2024 Annual Report" (2024) at 58. (Accessible here.)

United Kingdom and the EU in third place.[17] This concentration gives disproportionate influence to the countries in which major AI labs reside, as these labs are bound first by national regulation. It also suggests that the AI labs themselves have significant influence as, through self-regulation and other industry measures, they may act as de-facto regulators in this space, influencing the behaviour of the rest of the world.

- **Global power dynamics**: AI systems have clear benefits to national interests, including enhanced military capabilities, boosted economic advantage, and even enhancing a state's control over its people, which may be particularly appealing in authoritarian contexts. Thus, some states may aim to advance this technology as quickly as possible, and to prioritise their own national interests, rather than to support collaborative, globally harmonised regulatory frameworks.

- **Pre-existing regulatory cultures**: The process by which policies are made, as well as the institutional arrangements supporting different policies, vary across the world in line with the differing legal traditions and social priorities of different countries. The EU, for example, is known for its more precautionary approach and its comprehensive regulatory frameworks, such as the General Data Protection Regulation (GDPR), while the United States' approach is typically more sector-specific, federal, and tilted in favour of innovation instead of caution. As we shall see below, these pre-existing cultural differences represent different starting points from which states craft their AI regulation.[18]

- **Waves of public interest:** Since the release of ChatGPT – and its associated 'hype cycle' – there has been considerable appetite across the world for more comprehensive AI governance. Most of the CEOs of the major AI companies have supported calls for some kind of regulation, although the precise form they believe it should take remains ambiguous.[19]

- **Role of civil society and advocacy groups:** Efforts from international civil society and related groups have taken two broad themes – those primarily concerned with AI risks related to misinformation, prejudice, copyright ("AI ethics"),[20] as well as potential human rights violations, and those concerned with AI risks related to catastrophic harm ("AI safety").[21] These groups diverge in their beliefs about what is most important, although they overlap considerably in the steps they believe need to be taken to reduce risk from AI (for example, in their calls for AI models to be transparent, accountable, and subject to third-party safety audits). Both groups wield considerable influence, with the former being more

---

[17] See above n 16 at 61.
[18] A Engler "The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment" Brookings Institute (April 2023). (Accessible here.)
[19] See for example C Rozen "Regulate AI? Here's What That Might Mean in the US" The Washington Post (27 July 2023). (Accessible here.)
[20] This is typified by the November 2021 UNESCO "Recommendation on the Ethics of Artificial Intelligence" above n 1.
[21] A list of scientists, academics, policymakers, industry professionals, and other notable figures who hold this view can be found here.

influential among human rights and traditional civil society groups, and the latter carrying more weight amongst technical researchers and the AI labs themselves. The interests and arguments of these groups thus exert considerable influence on the regulatory environment.

Emerging AI governance approaches

As in many areas of technology regulation, the regulation of AI faces what is sometimes called the "Collingridge dilemma": if one creates regulation before the impacts of a specific type of technology are clear, that regulation may not function as intended (for example, it may stifle innovation without fully achieving its goals); but by the time its impacts are clear, the technology may be too entrenched to regulate effectively – or at least, its regulation will become more challenging over time, as the technology is already embedded in everyday life.[22] We have seen such dynamics at play in efforts to regulate social media over the last fifteen years. The challenge is even more acute in relation to AI, as the technology is advancing rapidly, and it is difficult to predict what capabilities will emerge from these systems over the next decade.

In recent years, countries across the world have been responding to this regulatory challenge in different ways. As regulation is rapidly evolving in this area, it can be useful to think about this by reference to the different types of regulatory efforts currently underway – only a small portion of which are binding.

There are a range of dimensions we can use to categorise these efforts:

- **Area of focus:** This relates to the specific aspects of an AI technology that the regulation seeks to address, such as:
    - Specific rights or harms: Regulation may seek to address particular AI risks, such as algorithmic bias or discrimination, or enforce particular principles or rights, such as transparency, oversight and accountability. As we will see, many regulations try to cut across these categories.
    - Innovation or development: Certain regulations may seek to create an enabling environment for AI development in their jurisdiction, such as by establishing institutions or funding mechanism to advance AI policy, research, and training.

- **Relevant parties:** Regulatory instruments on AI can confer rights, duties, or liability on different parties involved in the development, distribution, or use of AI technologies. These could include:
    - The *provider* of the technology (the person or entity who developed the AI or put it on the market);

[22] More information on the Collingridge Dilemma is available here.

- The *deployer* (a person or entity that is using AI technology developed by another party for some kind of official or non-personal use – such as a business or government);
- An end user or subject (a person who uses, or is subject to, an AI system developed and deployed by others);
- AI governance frameworks may impose different responsibilities on government actors and private industry.

- **Nature of framework**: The different governance frameworks in this area can be thought of on a spectrum from least to most binding, often reflecting how developed (or under-developed) the regulatory effort is in any given jurisdiction.

↤ Least binding			Most binding ↦
• World Economic Forum Framework for Developing National AI Strategy (2019)[23] • AU Data Policy Framework[24] • Voluntary commitments of private AI developers	• OECD AI Principles (2019)[25] • UNESCO Recommendation on the Ethics of AI[26] • UN resolution on safe, secure and trustworthy AI (2024)[27] • AU Continental AI Strategy[28]	• Blueprint for AI Bill of Rights (United States)[29] • National AI Strategy (United Kingdom)[30]	• AI Act (European Union)[31] • AI Convention (Council of Europe)[32] • Interim Measures for the Governance of AI (China)[33] • Various national data protection laws governing use of AI
Discussion documents and guidelines	**Principles and other soft law**	**National policy and strategy**	**National or regional laws**

[23] World Economic Forum "A Framework for Developing a National Artificial Intelligence Strategy" (August 2019). (Accessible here.)
[24] African Union Secretariat "AU Data Policy Framework" (2022). (Accessible here.)
[25] Organisation for Economic Cooperation and Development "Recommendation of the Council on Artificial Intelligence" (May 2019). (Accessible here.)
[26] Above n 1.
[27] Resolution on seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development GA Res 78/265 (2024). (Accessible here.)
[28] AU Continental Artificial Intelligence Strategy (2024). (Accessible here.)
[29] Executive Office of the President "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People" (October 2022). (Accessible here.)
[30] United Kingdom Department for Science, Innovation, and Technology "A pro-innovation approach to AI regulation" (2023). (Accessible here.) Note: This policy was issued under the Conservative-led government which was voted out of office in 2024. It is unclear how AI policy will change under the Labour Party administration.
[31] EU AI Act (2024). (Accessible here.)
[32] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (2024). (Accessible here.)
[33] Cyberspace Administration of China "Interim Measures for the Management of Generative Artificial Intelligence Services" (July 2023). (Accessible here.)

In terms of least binding instruments, we have **discussion documents** or "white papers", and voluntary commitments made by relevant stakeholders (notably AI companies) – which accordingly are unenforceable.

Next are **guidelines, declarations of principle, and related soft law instruments.** These are often made by stakeholders who do not necessarily have regulatory enforcement powers but can set the parameters for future regulation.

Further along are **national AI strategies** (which typically set out steps a country will follow to achieve its goals relating to AI, often including actions to expand development and research, and to produce policy and regulation) and **national AI policies** (which typically establish the principles which would guide a country's approach on AI).

National or regional AI legislation would be more binding, but the field is sparse: the first comprehensive AI law – the EU's AI Act – was only enacted in 2024, a few months before this toolkit was drafted.[34]

- **Sectoral scope:** Although some AI regulatory efforts may attempt to provide a comprehensive or universal framework (such as the EU's AI Act, discussed below), others may target specific sectors or uses, such as setting standards for government uses of AI (such a recent law enacted in the US state of New Hampshire), or setting rules for the use of AI in targeted contexts (such as recent laws enacted in the state of California, which regulate specific uses of AI in the entertainment industry and healthcare).

In the following section, we unpack more information about the emerging regulatory approaches in three major jurisdictions: the United States, China, and the European Union.

Regional examples

Given that regulation is quickly evolving across the world, this section looks at the high-level approach being taken by different significant global powers.

Europe

European Union's AI Act

In 2024, the EU passed the Artificial Intelligence Act (the AI Act), which is the first comprehensive legal framework for the development and use of AI systems. The AI Act takes a risk-based approach, categorising AI systems into several levels of risk, and creating different requirements and obligations for each one in their design or use within the EU.[35]

[34] European Commission "AI Act enters into force" (August 2024). (Accessible here.)
[35] The following analysis of the EU AI Act's risk categories draws on a guidance note by the European Commission Directorate-General for Communications Networks, Content and Technology, accessible here.

- **Unacceptable risk:** Certain systems are deemed unacceptably risky and are prohibited within the EU. These include systems that employ harmful manipulative 'subliminal techniques', systems used by public authorities for social scoring, and real-time biometric surveillance, such as facial recognition.

- **High risk:** AI systems are designated as 'high risk' if they are used as a safety component in products falling under the EU's health and safety legislation, or are used in certain specified areas such as education, migration, law enforcement, or the management of infrastructure.[36] These systems must be registered in an EU-wide database managed by the European Commission and must comply with a range of measures related to testing, data governance, transparency, human oversight, and cybersecurity.

- **Limited risk**: Systems that interact with humans (such as chatbots), as well as systems that generate audio, visual, and other types of content are designated as 'limited risk' and are only subject to basic transparency obligations (such as the requirement to disclose themselves to affected persons).

- **Low or minimal risk:** All other AI systems considered to pose low or minimal risk are not bound by any obligations, although the Act envisions the creation of codes of conduct to encourage the AI labs that develop them to voluntarily abide by the measures required of high-risk systems.

The Act provides for the establishment of the European Artificial Intelligence Office (the AI Office) which will oversee implementation and compliance, along with other oversight and advisory structures, and requires all EU member states to establish national authorities to oversee the Act's operation and enforcement domestically. These authorities would have access to confidential information (such as the source code of the relevant systems), and the power to impose fines for non-compliance.

The Act became law in August 2024, and various provisions will come into force in phases between 2024 and 2026.

During the law-making process, the Act faced opposition from global technology firms and other actors on the basis that it could jeopardise the EU's competitiveness and technological sovereignty without necessarily solving the relevant AI-related challenges.

[36] The full list of specified areas is as follows: Biometric identification and categorisation of natural persons; management and operation of critical infrastructure; education and vocational training; employment, worker management and access to self-employment; access to and enjoyment of essential private services and public services and benefits; law enforcement; migration, asylum, and border control management; assistance in legal interpretation and application of the law. See reference above.

Human rights groups, on the other hand, have criticised the AI Act for far-reaching exemptions and exceptions which may enable human rights violations through the use of AI.[37] For example, the Act's safeguards do not apply to AI systems used for national security[38] and critics argued that the Act's partial ban on *real-time* biometric surveillance such as facial recognition should have applied to all forms of biometric surveillance.[39] (It is unclear, for example, that 'retrospective' use of AI for biometric surveillance, such as conducting facial recognition on video footage which was recorded last week, is necessarily less invasive than doing it on live video footage.)

Council of Europe's AI Convention

In May 2024, the Council of Europe adopted the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (the AI Convention),[40] described as the first internationally binding treaty on AI.

The AI Convention's broad aim is to oblige all signatory states to adopt laws and administrative measures which ensure that the full lifecycle of any AI system is consistent with human rights, democracy, and the rule of law.[41]

- **Scope:** These measures apply principally to the use of AI by public actors *or* private actors serving a public function, with less binding requirements for the design or use of AI solely by private actors.

- **Principles:** State parties are obliged to enact legislation and other measures to ensure that the full lifecycle of such AI systems are consistent with their existing obligations to protect human rights, democratic process, and the rule of law, and conform to common principles for AI governance.[42]

- **Obligations:** States are also obliged to provide for reasonable remedies and complaints mechanisms in the event that an AI system is inconsistent with these principles,[43] and adopt measures for AI impact assessments or to consider bans on certain AI systems that are inconsistent with these principles.[44]

---

[37] Leufer "Why human rights must be at the core of AI governance" Access Now (September 2024). (Accessible here.)
[38] European Center for Not-for-Profit Law "Packed with loopholes: why the AI Act fails to protect civic space and the rule of law" (April 2024). (Accessible here.)
[39] Article 19 "EU: AI Act fails to set gold standard for human rights" (April 2024). (Accessible here.)
[40] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law CETS No. 225 (2024). (Accessible here.)
[41] Ibid, Article 1.
[42] These principles, outlined in Articles 7-14, are: Human dignity and individual autonomy; Transparency and oversight; Accountability and responsibility; Equality and non-discrimination; Privacy and personal data protection; Reliability; and Safe innovation.
[43] Ibid, Article 14.
[44] Ibid, Article 15.

The AI Convention will come into force once it has been ratified by five states, including at least three member states of the Council of Europe. No states had ratified at the time of this publication, but the Convention had received non-binding signatures from eight member states, including Norway and the UK, as well as from the US, Israel, and the EU.[45]

Human rights groups have criticised the AI Convention on a range of grounds, including its broad exemptions for AI systems used for national security and defence purposes, and for applying lower standards and protections to AI systems in the private sector.[46]

United States

In contrast to the EU's comprehensive and centralised approach, the United States' approach has been more piecemeal, sector-specific, and distributed across various federal agencies,[47] although US state legislatures produced a rush of non-federal AI regulation in 2024. The federal government has primarily led with non-binding interventions and sector-specific guidelines which fall within the powers of the executive. For example, the US Copyright Office has issued rulings that suggest that most text, images, and videos created by AI systems cannot be copyrighted as original works. The absence of binding federal regulation could be read as either a policy orientation, or a function of the polarised and partisan make-up of the US congress, which has a poor record on tech regulation (the US still has no federal data protection law, for example) and which was historically unproductive in 2023.[48] There have been proposals for an Algorithmic Accountability Act, which would require companies to evaluate the bias and effectiveness of their AI systems, and would require the country's Federal Trade Commission to enforce this requirement,[49] but at the time this toolkit was produced there were no signs of progress. As we will explore below, in the absence of binding federal action, many US states have undertaken their own policymaking at state level.

Administrative actions

- **Executive order:** In October 2023 US President Joe Biden issued an Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence,[50] which mandated a range of federal agencies to develop standards for safe and ethical design and use of AI in their various sectors, and which placed new reporting requirements for companies developing AI with national security implications to share their testing data with the US government. Although these actions were largely directed at internal processes of the federal government, within nine months agencies had undertaken over a hundred actions or policy processes in

[45] A schedule of non-binding signatures and ratifications for the AI Convention is accessible here.
[46] See European Network of National Human Rights Institutions "Statement of Concern on Draft Convention on AI" (March 2024). (Accessible here.)
[47] Above n 18.
[48] C Hunt "Is this the least productive congress ever?" The Conversation (August 2024). (Accessible here.)
[49] K Piper "There are two factions working to prevent AI dangers. Here's why they're deeply divided" Vox (August 2022). (Accessible here.)
[50] Executive Office of the President "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence" (October 2023). (Accessible here.)

response.[51] This suggests more meaningful progress than that of a 2019 Executive Order from the Trump administration ("Maintaining American Leadership in Artificial Intelligence") which also mandated various federal agencies to develop plans to regulate AI applications; by December 2022, only one of the 41 major agencies (the Department of Health and Human Services) had meaningfully created such a plan.[52]

- **Blueprint for AI Bill of Rights:** Prior to its Executive Order, the Biden administration also issued the 2022 'Blueprint for an AI Bill of Rights',[53] a non-binding document that sets out five principles[54] and associated practices to guide the development and use of AI. It tasks different federal agencies with implementation in their respective policy sectors (like health, labour, and education). The Biden administration also secured voluntary commitments from major AI developers in the US to meet certain standards for testing and transparency of their systems; by mid-2024 16 companies had signed on to these commitments, including Amazon, Anthropic, Apple, Google, Inflection, Meta, Microsoft, and OpenAI.

The re-election of Donald Trump to the US Presidency in 2024 may result in these administrative actions being rolled back, and federal regulation on AI seems unlikely in the foreseeable future. In the absence of binding policy at the federal level, many state legislatures across the United States have introduced, and in some cases enacted, state laws relating to AI.

California state law

In particular, in 2024 the state of California enacted a suite of *17* laws addressing various AI-related concerns.[55] Although the Governor vetoed the most far-reaching AI bill produced by California lawmakers (see below) the remaining bills comprise the most substantive AI regulation in US state law, both because of their content, and because of California's outsized regulatory influence as the biggest state economy in the US, and as the home state of many of the world's leading AI developers.

- **Vetoed bill on AI safety:** California's governor vetoed the most prominent AI bill sent by state legislators, the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, in September 2024. If it had been enacted, the Act would have put stringent new requirements for developers to prevent their technologies from causing serious harm,

[51] Executive Office of the President "FACT SHEET: Biden-Harris Administration Announces New AI Actions and Receives Additional Major Voluntary Commitment on AI" (July 2024). (Accessible here.)
[52] Above n 18.
[53] Executive Office of the President "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People" (October 2022). (Accessible here.)
[54] The principles include: safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation an human alternatives, consideration, and fall back. More about these can be accessed here.
[55] Office of the Governor of California "Governor Newsom announces new initiatives to advance safe and responsible AI, protect Californians" (September 2024). (Accessible here.)

including imposing severe fines and civil liability for failing to comply with mandatory reporting and safety procedures.[56]

- **Other California AI bills:** The governor did sign into law 17 other bills addressing a raft of other issues in AI, most of which come into force in 2025 and 2026.[57] Their provisions include:

    - Expansions of the state's data privacy law to ensure they extend to personal information used in the training or operation of AI tools;
    - Stronger protections for actors and other entertainers against AI-generated replicas of their voice or likeness;
    - Requirements for major generative AI systems to make AI-generated content easier to identify, through 'watermarking', providing tools to verify if content was created or modified with AI, and other measures;
    - Minimum transparency requirements for generative AI systems to disclose high-level information about the data used to train them;
    - A requirement for health providers to provider a disclosure and disclaimer when using generative AI to communicate clinical information to patients, and to provide clear instructions for how to contact a human healthcare provider;
    - Expansions of the state's laws on child sexual abuse material (CSAM) and non-consensual intimate image (CSII) to ensure their penalties extend to media generated or modified with AI.

Colorado state law

In May 2024, the state of Colorado enacted Senate Bill 24-205 (the Colorado Artificial Intelligence Act), which approaches AI safety through the lens of consumer protection, and which applies to anyone doing business in the state. A summary of its provisions is as follows:[58]

- The Act principally applies to "high risk" AI systems, defined as those whose outputs result in a "consequential decision" – meaning a decision which materially affects a person's access to education or employment opportunities, financial services, housing, healthcare, or essential government services – or the cost thereof.

- It creates obligations on both developers of AI technologies (those who design the technology), and deployers (those who use the technology for a commercial purpose);

- Developers of such models need to take proactive measures in their design process to address risks of algorithmic discrimination resulting from their technology, and to disclose

[56] See Anderson and others "Raft of California AI Legislation Adds to Growing Patchwork of US Regulation" whitecase.com (October 2024). (Accessible here.)
[57] Ibid.
[58] See K Anderson "The Colorado AI Act: What you need to know" bakertilley.com (September 2024). (Accessible here.)

information about its functioning, testing, and safeguards against discrimination, and high level information about its training data;

- Deployers must take similar steps to document and disclose such details, and most report any incidents of discriminatory outcomes to the state Attorney General;

- Where such an AI system makes or contributes to a consequential decision that is adverse to a consumer, the deployer must provide the consumer with reasons for the decision, including an explanation of how the AI system contributed to the decision, and what data was used to reach it. The deployer must also give the consumer an opportunity to appeal the decision and get a human review.

Other US state laws

According to the National Conference of State Legislatures, as of September 2024 at least 44 other US states introduced AI bills in 2024, and at least 30 of those states enacted one or more AI laws.[59] Of these:

- At least eight regulate the use of AI in the context of elections or political advertising,[60] for example by requiring a disclosure for political advertising that uses AI-generated content, or by making it an offence to distribute deceptive AI-generated media (such as 'deepfakes') to influence an election;

- At least four expand the scope of existing laws on child sexual abuse material (CSAM) and non-consensual intimate image (CSII) to include media generated or modified with AI;[61]

- At least four created some prohibition on certain uses of AI by government bodies.[62] For example, New Hampshire's HB 1688 prohibits state agencies from using AI for real-time biometric surveillance, except with a warrant, or from categorising people by behaviour or class if this results in unlawful discrimination; Utah passed criminal justice amendments which restrict the court' reliance on algorithmic assessments in probation decisions;

- At least seven states passed laws or resolutions which do not directly regulate the design or use of AI, but serve primarily to establish an advisory body or task force to develop recommendations or policies to guide the state's approach on artificial intelligence.[63]

As of September 2024, more than 200 other AI bills were listed as 'pending' in state legislatures.[64]

---

[59] See database published by NCSL "Summary: Artificial Intelligence 2024 Legislation" (September 2024). (Accessible here.)
[60] Alabama, Colorado, Florida, Florida, New Hampshire, New Mexico, Oregon, Utah, and Wisconsin.
[61] Alabama, Florida, North Carolina, and South Dakota.
[62] Maryland, New Hampshire, Utah, and New York (pending enactment by the governor).
[63] Delaware, Florida, Indiana, Massachusetts, Pennsylvania, Tennessee, and West Virginia.
[64] See above n 59.

China

In July 2023, China published its "Interim Measures for the Management of Generative AI Services", which provide a window into the country's approach to AI regulation.[65] These measures hit many of the same topics as in the rest of world (discussed further below), although they are notable for their authoritarian slant. Thus, in addition to discussions on the need to promote fairness, transparency, and international cooperation, one finds the following clause:[66]

> "The provision and use of generative artificial intelligence services shall abide by laws and administrative regulations, respect social morality and ethics, and abide by the following provisions: (1) Adhere to the core values of socialism, and must not generate incitement to subvert state power, overthrow the socialist system, endanger national security and interests, damage national image, incite secession, undermine national unity and social stability, promote terrorism, extremism…"

AI governance in China is led by the Cyberspace Administration of China, although this may change as the scope of AI expands.[67] The public/private distinction is also much less clear in China, as the Chinese government plays a much more prominent role in China's AI ecosystem[68]

Other notable Chinese AI regulations include its 2021 Regulation on Recommendation Algorithms and its 2022 Rules for Deep Synthesis (synthetically generated content).[69]

Trends and themes

As we have seen above, there is considerable variation in how countries intend to implement their various commitments in relation to AI, and it is unclear what policy approach most other countries will take. Yet as we unpack below, across the emerging approaches there is surprisingly consensus on many of the major issues. Transparency, accountability, data protection and other principles are all regarded as important principles across these frameworks, despite enormous differences in the regulatory cultures in each jurisdiction.

AI impact assessments

What is an AI impact assessment?

---

[65] Cyberspace Administration of China "Interim Measures for the Management of Generative Artificial Intelligence Services" (July 2023). (Accessible here.)
[66] Ibid, at article 4(1).
[67] M Sheehan "China's AI Regulations and How They Get Made" Carnegie Endowment for International Peace (July 2023). (Accessible here.)
[68] J Ding and J Xiao "Recent Trends in China's Large Language Model Landscape" Centre for the Governance of AI (April 2023) at 3. (Accessible here.)
[69] M Sheehan "China's AI Regulations and How They Get Made" *Carnegie Endowment for International Peace* (July 2023) at 4 (Accessible here.)

An impact assessment is a process to identify and mitigate potential harms that could result from a particular project, tool, or policy. Outside of the context of AI, for example, *environmental* impact assessments are often a legal requirement before undertaking a major construction project; more recently, *data protection* impact assessments have become a common requirement for the design of policies, technology, or practices, which may involve the use of people's person information.

In the context of AI, impact assessments are widely held as an important process to mitigate the potential harms which could result from a particular AI technology. They are generally focused on human rights implications, but could include environmental or other concerns as well.

Depending on the regulation and risk, an impact assessment could be undertaken at different stages of the process – and more than once. For example, it may be conducted before or during the design and development of the technology *(ex-ante)*, or after it is completed *(ex-post)*. It may be a requirement of the developers of the technology (to test and refine their design), or of the deployers of the technology (to test and refine their use of the technology) – or both.

If the assessment identifies a potential harm that outweighs any benefit that comes from the technology, it could result in the technology or use being re-designed, or even discontinued.

Importantly, impact assessments are necessarily grounded in the context of the technology or use that is being assessed, which means there is not a single format for an AI impact assessment.[70]

AI impact assessments feature prominently as a safety procedure in guidelines and standards for AI governance. The two preeminent AI laws enacted in 2024, the EU AI Act and Colorado's AI Act, both include requirements for developers and/or deployers to conduct AI Impact Assessments for 'high risk' systems.

AI and labour rights

What is the intersection between AI and labour rights?

Much has been said about the impact of AI and automation on the job market in two overarching ways. First, AI's capacity to replace jobs that have been traditionally conducted by humans. Second, AI's capacity to create new jobs and skills. It therefore becomes important for regulatory frameworks to consider and protect workers' rights.

The UNESCO Recommendation on AI Ethics calls on Member States to consider labour rights by using impact statements to understand the rights implications of AI. The Recommendation puts forward five suggestions on this question. These may be summed up as follows:[71]

[70] For a compilation of resources on AI impact assessments, visit https://ai.altadvisory.africa/impact/.
[71] Above n 1 at 36.

- A potential revision of education requirements across all levels can equip people to either integrate into or adapt to a job-market that uses AI systems. Transparency around these skills is essential.

- The development of bridging courses across all industries, and enterprises of various sizes, can help boost the skill set needed by workers. These courses should be done through collaboration between government, academic institutions, industry, workers' organisations, and civil society.

- Member States must proactively ensure that at-risk employees undergo a fair transition and that "safety net" programmes are created for those who may not be retained in the job market. Revisions to applicable tax regimes may be a useful measure in response to mass unemployment as a result of AI-based automation.

- Member States should encourage ongoing research into the impact of AI systems on the labour environment and forecast trends and challenges.

- Lastly, competition and consumer protection frameworks should be bolstered to mitigate abuse of domain market positions concerning the lifecycle of AI systems.

Closing commentary

It is interesting to note how existing public interest concepts like transparency and accountability are being adapted to support the governance of AI. Readers should consider to what extent these adaptions have been successful, and to what extent there are gaps in current frameworks. Although there are already an abundance of frameworks, declarations, and related regulatory instruments, we are still in the very early days of AI governance. New regulations are being developed and propagated monthly, often displacing, or overwriting what came before.

A close look at existing documents reveals that while there is relative consensus on which values and principles are most important in this space, there is considerable divergence on how those values and principles should be operationalised – with Europe, America, and China all taking distinct approaches. It remains to be seen which approach is most effective – and indeed, the most appropriate approach for a given region may depend on its local context.

In the face of all this regulatory uncertainty and development, there is a huge opportunity for civil society and journalists to exert influence on the future of AI governance. Given the complexity of this space, it is thus also important for civil society and related actors to understand what outcomes most desirable, and what methods might achieve them.

Part 2 of this toolkit will turn to the emerging approaches to AI governance in Africa.

Ends.