Governing GenAI *for all* and *for good*

Gabriela Ramos*

Abstract: The advent of Artificial Intelligence and, more specifically, Generative AI, promises immense potential to revolutionise various sectors and improve human life. However, the rapid development and deployment of AI also raise critical societal concerns. By prioritising human values over short-term gains, the paper sustains that we can shape the direction of AI and ensure a positive impact on economies and societies worldwide. It calls for a proactive approach to AI governance, as opposed to self-regulation or ex-post measures. It underscores the importance of implementing guardrails to avoid amplifying existing inequalities and potential new challenges. It presents the UNESCO Recommendation on the Ethics of AI as a normative framework for AI governance worldwide. It also recalls that the Readiness Assessment Methodology (RAM) can help countries discuss specific advice on the kinds of rules, institutions and policies they may use for ethical AI development. Several countries have already started implementing this framework and methodology to ensure AI respects human rights, benefits all humanity, and follows ethical principles.

Keywords: Artificial Intelligence (AI), Generative AI (GenAI), Foundation Models and Private Sector, inequalities, markets concentration and ethics, Multilateral cooperation; UNESCO Recommendation on AI & Ethics

Life in the Age of Artificial Intelligence

We have officially entered the Age of Artificial Intelligence. AI has the potential to make our lives easier and smoother. It can help us find a home, get information, improve our finances, look for and get a job, or do our taxes. It can improve food production and management by making agriculture more efficient and increasing food safety. It can help predict and combat natural disasters and preserve biodiversity by developing energy-efficient cities and better power storage and distribution systems. It can increase the accuracy of medical diagnoses when combined with doctors' assessments. And it can make us more productive by saving us resources for routine work and leaving us more time for creative thinking. Generative

* Assistant Director-General for Social and Human Sciences of UNESCO. Views are personal.

AI (GenAI), the most promising and powerful new trend in technology, could bring in $2.6 trillion to $4.4 trillion annually across all industries, equivalent to the entire GDP of the UK (Ramos, 2023; McKinsey, 2023).

At the same time, without clear guardrails, AI can reproduce and amplify many of the social challenges we face and even create new inequalities. In 2023, half of the world's population still lacks adequate internet access, and the gap between regions, gender, income, language and age groups remains. Nearly 90 per cent of people in Europe are online, but only 21 per cent of women in low-income countries use the internet. Digitally deliverable services account for almost two-thirds of global services trade, but significant digital learning gaps are emerging and growing, necessitating a global digital pact (UN, 2023).

Market concentration is also a major problem, as AI developments and the underlying data are in the hands of a few companies, with only two countries – the United States and China – owning most of the innovations. It is true that some private companies are proactively putting principles and processes in place to make AI sustainable and ethical, but this is not the majority, and business interests and geopolitical considerations – which are often short-sighted - prevail. As US President Biden underlined recently, these principles are still voluntary. Non-diverse AI teams, non-representative databases and opaque and biased algorithms can do harm, especially to those already at risk, be they businesses or individuals, children and young people, women or even democracies. AI has become the latest technological competition (after 5G, semiconductors, platforms, pharma) driven by short-term benefits or geopolitical considerations at the expense of global interests (Maslej et al., 2023).

Not to mention that authorities in some countries are already using AI-powered facial recognition technologies to monitor political dissidents and conduct mass surveillance, which weakens democracies.

Given the above contrasting trends, Who is in charge here? Too many important technologies are currently being used without clear guardrails. Of course, it would not be the first time that regulators have lagged behind market developments. In the case of AI, however, the gap is significant. There is a need for a quantum leap in the governance of technologies for building the capacity to understand them and enforce ethical guardrails. (Mazzucato and Ramos, 2022; UN, 2023). We need to create rules and institutions that are guided by our values as a society. This is not about technology. It is about the governance frameworks shaping their development and use, and the ethical and moral values that sustain them.

AI Policy Agenda for the Common Good: Institutions, Rules and Skills

For once, we should align technological developments with our overall goals. With many ministries within governments having something to say, we need more coordination and even innovative governance. Establishing

high-level offices to coordinate digital and artificial intelligence strategies is an international good practice that ensures coherence and political participation at the highest level. But this is not enough. AI should be a State – not a government - policy. Empowering data protection authorities or strengthening or building new independent digital economic agencies that assess and licence AI developments (similar to the US Food and Drug Administration) should be seriously considered. Moreover, the policy agenda for AI in the XXI century should couple the institutional and regulation agenda with substantial investment in skills, both in governments and in the labour markets to shape the technological revolution for the common good (Mazzucato and Ramos, 2022).

The UNESCO Recommendation on the Ethics of AI, a normative framework adopted by 193 member states, contains the most detailed action plan currently available at the international level to build these pillars. The Recommendation's aim is to promote and protect human rights, human dignity, environmental sustainability and gender equality. It promotes principles such as accountability, transparency and, above all, the rule of law. It contains concrete policy chapters calling for better data governance and sets out policy tools to bring about policy change. These include the Ethical Impact Assessment, which identifies the impact of AI systems both ex-ante and ex-post, and the Readiness Assessment Methodology (RAM), which helps countries assess their readiness and forms the basis for UNESCO policy and capacity-building support. As of June 2023, nearly 30 countries, and more to come, have begun using the Recommendation to enact national laws that ensure AI respects fundamental freedoms and human rights and benefits all humanity. India will soon pilot RAM.

UNESCO's efforts are gaining visibility and impact, and key global actors are now taking responsibility for refocusing the debate. Perhaps most striking is the return of the US to UNESCO, recognising the relevance of UNESCO's mandate and in particular its leadership role in AI ethics policy. As addressed by Secretary Blinken at the Congress: "It was important (for the US) to be at the table that is defining the rules for AI … things that are happening at UNESCO actually matter. They are working on rules, norms and standards for artificial intelligence. We want to be there". This policy momentum is reinforced by the European Union developing a holistic regulatory framework and a forthcoming AI Act as well as supporting the implementation of UNESCO's work on ethics in developing countries. It is also particularly important to mention the AI strategies of key emerging economies such as Brazil or India and India's initiatives at the Presidency of the G20.

The era of light-touch ex-post self-regulation is ending for good with the arrival of big language models and GenAI. The impact of data bias has been magnified and previous AI developments pale in comparison (e.g. when limited data samples were used to select a

company's CEO or design a healthcare device). Privacy concerns (when applied to media use) have come to the fore. As The Economist recently noted, GenAI developers themselves have often been surprised by the power of their creation. The bottom line is that AI and generative AI, like any other major product, should be tested for safety and trustworthiness before they are launched into the market. UNESCO highlights the need for this impact assessment to be done in advance.

Foundation Models: Special Features and Initiatives of the Private Sector

Since mid-2022, the release of AI foundation models for text, images and audio files to the public and the massive growth of their user base has intensified and broadened the debate about the risks they pose to labour, education, scientific research and democracy, as well as their potential negative impact on cultural diversity and cross-cultural interactions. Foundation models are AI systems characterised by the use of large machine learning models trained on massive unlabelled datasets, using significant computational resources. Examples include large language models (LLMs) such as the GPT series and Bard, and image generator tools such as DALLE 2 and Stable Diffusion (UNESCO, 2023a).

The speed of their adoption and their increasing capacities, measured in weeks rather than years, deepen the known risks of AI. Preliminary assessments based on UNESCO Recommendation confirms that LLMs can provide misleading, inaccurate or false information without making this clear to the user (ChatGPT has only recently introduced a disclaimer). Their impact on science, research, education and work is also magnified by the range of tasks the tool can perform. This makes the list of unknowns even longer and the risks in human-machine interaction greater. Generative AI can shape people's minds, thoughts and behaviour. Implementing UNESCO's ethical framework is therefore a must. The fact that these AI models are often described by their developers as "experimental" and issues often only come to light after they have been released to the public also highlights the need for ex-ante regulation and the establishment of liability regimes.

It should be acknowledged that some leading private companies and business associations have launched initiatives to guardrail foundation models' functioning and use. For example, NASSCOM, the National Association of Software and Service Companies of India, has published its guidelines for the 'responsible' use of generative AI and adopted the UNESCO Recommendation on AI & Ethics (NASSCOM, 2023). As a step forward and in a similar spirit, Microsoft recently presented its white paper on AI governance, proposing specific regulations for each of the AI technology layers. They proposed applying existing legal protections at the applications layer to the use of AI (e.g. banking, insurance, commerce); adding new AI expertise and capabilities to existing regulatory agencies; and developing new laws and regulations for these AI foundation models; even

establishing a licencing regime similar to that for telecommunications network operators and critical infrastructure providers (MICROSOFT, 2023). Microsoft is also leading the Business Council, which was established to implement the UNESCO standard in a multistakeholder manner.

Given that so much knowledge and experience exist in the private sector, synergies between the public and private sectors are both inevitable and desirable. To take this work forward, UNESCO relies on a large group of partners in the public and private sectors, as well as in civil society, to ensure that the Recommendation is translated into concrete policy action and regulatory insights. With the support of the Japanese government, the Patrick McGovern Foundation, the European Commission and CAF, and the Development Bank of Latin America, UNESCO is now deploying its tools in a large group of countries and has established the AI Experts without Borders and Women4EthicalAI networks. UNESCO is also working with a large number of knowledge institutions and will be establishing the Observatory of Ethics of AI with the Alan Turing Institute.

These examples are steps in the right direction, but they are not enough. Bold public action is urgently needed.

UNESCO Recommendation on AI & Ethics in Action: Governing GenAI for Good and for All

UNESCO recently analysed foundation models and GenAI through the lens of its Recommendation on the Ethics of Artificial Intelligence and concluded that three challenges stand out: Fairness and verifiability, labour market impacts and environmental sustainability (UNESCO, 2023a).

Quality of data, fairness and verifiability

The quality, coverage transparency and verifiability of data are particularly important to prevent disinformation and misinformation and to address discrimination in ChatGPT and models alike. These models tend to be opaque, both in terms of the dataset used to train them (and some even refuse to disclose what data was used). Providing transparency and explainability should involve, at least, providing real references for the factual claims made, so that users can understand where the answers they are getting come from, and are better empowered to judge their truth, bias, and trustworthiness – while also, where relevant, giving credit to the creators of the content from which the tool is deriving its outputs (UNESCO, 2023a). The principle of fairness and non-discrimination places particular emphasis on the inclusion of all members of society, especially persons with disabilities, women and children, and all marginalised groups, taking into account their specific needs and language requirements. A particular focus should be placed on gender inequalities and biases.

Labour markets and skills

For millions of low-income people with good connectivity, the democratising

effect of access to knowledge and digital services that ChatGPT brings is excellent, especially for facilitating autonomous learning or breaking down barriers to accessing research assistance. But as long as the unconnected half of the world's population is not able to access these services, gaps will remain and grow.

In addition, foundation models reinforce concerns about the impact of AI on labour markets and the speed and depth with which certain jobs will change; in particular tasks such as reasoning, writing, creating graphs and analysing data that differ from software and robots, which impact low - and middle-skilled tasks (Webb, 2020). Not to mention that foundation models' training is labour intensive, often using 'ghost workers', with sub-optimal working conditions, who provide human feedback to optimise reinforcement learning, oftentimes from low-income countries.

We need calls for an ambitious forward-looking agenda for cognitive and socio-emotional skills with a focus on communication, problem solving, creativity and teamwork. A significant number of socio-emotional skills, combined with more cognitive skills, seem to be in constant demand (Ramos, 2022). Moreover, retraining and upskilling are key. Analyses of the workers' performance on the job and on workers' skills endowment and learning opportunities, show that the cost of changing jobs (either within the same firm or elsewhere) is not trivial (Andrieu et al., 2019). This requires coordinated action by government, companies, trade unions, civil society and workers to "put in place" upskilling and reskilling programmes, find effective mechanisms to retain workers during these transition periods, and explore 'safety net' programmes for those who cannot be retrained (UNESCO, 2023a). Job analysts have mainly focused on assessing how jobs will change and understanding the jobs of the future, but the real challenge is to provide the necessary support for the transition period, which can be long and painful.

A discussion is taking place under the Indian Presidency of the G20 Digital Economy Working Group (DEWG), in particular in Priority 3 "Digital Skilling", where UNESCO is a knowledge partner. The aim is to help create a future-ready workforce. The starting point is to recognise the skills gaps that characterise the economies and societies and how disruptive and divisive these can be. A G20 Toolkit for the Design and the introduction of digital upskilling and reskilling programmes is being developed, to identify good practices and help G20 members better assess and improve their skills strategies. The work of the G20 DEWG under the Indian Presidency has also highlighted the need for a widespread measurement of skills, capabilities and competencies to enable cross-country comparisons and create a common understanding of digital skills across borders. This would facilitate tapping into a global talent pool and help address supply and demand gaps of human capital faced by economic actors and workers, and facilitate working with AI (Samek and Squicciarini, 2023).

Environmental sustainability

The environmental footprint of the large foundation models is considerable. For example, the training run of BLOOM (the less energy-intensive of the four language models analysed by the Stanford Institute for Human-Centered AI) emitted 25 times the amount of a round-trip passenger flight from New York to San Francisco and consumed as much energy as an average US household in 41 years (Maslej et al, 2023).

This calls for assessing the direct and indirect environmental impact throughout the AI system life cycle, including, its carbon footprint, energy consumption and the environmental impact of raw material extraction for supporting the manufacturing of AI technologies, and reducing the environmental impact of AI systems and data infrastructures. Moreover, when choosing AI methods, given the potential data-intensive or resource-intensive character of some of them and the respective impact on the environment, Member States should ensure that AI actors, in line with the principle of proportionality, favour data, energy and resource-efficient AI methods (UNESCO, 2021).

Fighting 'AI pauses' and regulation myths

Following communiques and open letters from tech industry leaders calling for a pause in the training of the most powerful AI systems, UNESCO urged countries to fully implement its Recommendation on the Ethics of Artificial Intelligence without delay. We do not need a pause; we need to redouble efforts to ensure that governments are able to shape inclusive and fair technological development. UNESCO's global normative framework provides the necessary guarantees for cutting-edge foundation and GenAI models. As industry self-regulation is not sufficient, the Recommendation are tools to ensure that AI developments follow rule-of-law principles, that harm is avoided, and that, when harm is done, accountability and redress mechanisms are available for those affected (UNESCO, 2023b).

As we said, more than 30 countries in all regions of the world are already working with UNESCO to develop AI checks and balances at the national level. They draw on the Recommendation and on the Readiness Assessment Methodology (RAM) to receive advice on the kinds of rules and policies needed for the ethical development and use of AI, but also to explore institutional innovations that could take on the oversight role in AI. UNESCO calls on all countries to join the movement which is leading to build ethical AI. A progress report will be presented at the UNESCO Global Forum on AI Ethics in Slovenia in early 2024.

Despite the oft-cited myth that AI cannot be regulated, there is growing evidence to the contrary. Claiming that AI is so dynamic and ubiquitous that it cannot be regulated, and refusing to train large language models because they could never be implemented (because if you prevent large companies from doing so, others will fill the gap), is the same as claims about chemical or biological weapons. And regulation has generally

worked for them. Another argument is that governments do not have the expertise or human capital to regulate AI. But this is precisely why it is time for government to invest much more in attracting and re-skilling internal talent, as one of the pillars of public policy (Acemoglu, 2023). The new or strengthened AI supervisory and regulatory body is central to advancing the regulatory agenda. This reinforces the belief that sound AI governance and regulation is indeed good for good businesses, similar to how companies that show improvement in environmental, social and governance values tend to have higher shareholder returns compared to industry peers in the period following improvement in social responsibility scores (McKinsey, 2022). In some countries, such as Spain, responsible AI labelling is being introduced in line with proposals to extend social responsibility to digital issues ('EDSG' as proposed by Benjamins and Melguizo, 2022).

We have the tools and the political will, let us move forward!

The future of our society is at stake. Not only do we need to solve the problems and control the risks of AI, but above all we need to shape the direction of digital transformation and technological innovation more broadly. AI, especially GenAI, is becoming a general-purpose technology in the midst of a global economic and geopolitical race that impacts multiple/all industries and societies. It therefore needs context, i.e. history and values (Kissinger, Schmidt and Huttenlocher, 2021). ChatGPTs and LLMs, in particular, raise high expectations of their service offering. These could be substantial, accelerating productivity and income, well-being and inclusion. However, their widespread use also highlights the risks associated with the way these technologies are currently deployed, responding to a frantic technological race between economic actors and countries, rather than serving the public good. To get this right, we need the right institutions and policy frameworks, and that is what UNESCO has been mandated to do.

The ethical approach to AI led by the UNESCO Recommendation on the Ethics of Artificial Intelligence can deliver fair, sustainable and inclusive outcomes. But it cannot do so without capable governments that protect the rule of law online, State structures that are adapted to the new AI world, and private developers who are accountable for putting people – not profits or geopolitical considerations – first. Only then can the Age of (Gen)AI bring the progress we hope for. We hope that the concerns raised by these technologies will help us build more solid governance frameworks that will have a positive impact on our economies and societies. That is why initiatives like India's under its G20 Presidency to strengthen human capital for the digital and AI age and their commitment to implementing the UNESCO Recommendation on the ethics of Artificial Intelligence are key.

References

Acemoglu, D. 2023. https://threadreaderapp.com/thread/1667509643076091904.html, MIT.

Andrieu, E., S. Jamet, L. Marcolin and M. Squicciarini. 2019. 'Occupational transitions: The cost of moving to a "safe haven"', OECD Science, Technology and Industry Policy Papers, 61, OECD Publishing, Paris, https://doi.org/10.1787/6d3f9bff-en.

Benjamins, R. and A. Melguizo. 2021. 'Sin digitalización no hay sostenibilidad', El País. https://elpais.com/america-futura/2022-09-06/sin-digitalizacion-no-hay-sostenibilidad.html

Kissinger, H.A., E. Schmidt and D. Huttenlocher. 2021. 'The Age of AI: And our Human Future.' Little, Brown and Company, NY. https://www.hachettebookgroup.com/titles/henry-a-kissinger/the-age-of-ai/9780316273800/?lens=little-brown

Maslej, N., M., L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J.C. Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark and R. Perrault. 2023. The AI Index 2023 Annual Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. https://aiindex.stanford.edu/report/

Mazzucato, M. and G. Ramos. 2023. 'AI in the Common Interest', Project Syndicate, December. https://www.project-syndicate.org/commentary/ethical-ai-requires-state-regulatory-frameworks-capacity-building-by-gabriela-ramos-and-mariana-mazzucato-2022-12

McKinsey. 2022. 'Does ESG really matter – and why?', McKinsey Quarterly, August. https://www.mckinsey.com/capabilities/sustainability/our-insights/does-esg-really-matter-and-why

McKinsey. 2023. The economic potential of generative AI: The next productivity frontier. McKinsey Digital, June. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

MICROSOFT. 2023. 'Governing AI. A Blueprint for the Future'. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

NASSCOM. 2023. Guidelines for Generative AI, June.https://nasscom.in/ai/responsibleai/images/GenAI-Guidelines-June2023.pdf

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue.https://openai.com/blog/chatgpt/

Ramos, G. 2022. "A.I.'s Impact on Jobs, Skills, and the Future of Work: The UNESCO Perspective on Key Policy Issues and the Ethical Debate," New England Journal of Public Policy, 34(1). Available at: https://scholarworks.umb.edu/nejpp/vol34/iss1/3

Ramos, G. 2023. 'What Lies Beyond the AI Tipping Point?', Project Syndicate, February. https://www.project-syndicate.org/onpoint/ps-commentators-respond-what-lies-beyond-the-ai-tipping-point?barrier=accesspaylog

Samek L and M. Squicciarini. 2023. 'AI Human Capital, Jobs and Skills', in Appio, F.P., La Torre, D., Lazzeri, F., Masri, H., & Schiavone, F. (Eds.), Impact of Artificial Intelligence in Business and Society: Opportunities and Challenges (1st ed.). Routledge. https://doi.org/10.4324/9781003304616

UN. 2023. 'A Global Digital Compact - an Open, Free and Secure Digital Future for All', Our Common Agenda Policy Brief 5, May. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-gobal-digi-compact-en.pdf

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

UNESCO. 2023a. Foundation models such as ChatGPT through the prism of the UNESCO Recommendation on the Ethics of Artificial Intelligence, SHS/2023/PI/H/12. https://unesdoc.unesco.org/ark:/48223/pf0000385629

UNESCO. 2023b. 'Artificial Intelligence: UNESCO calls on all Governments to implement Global Ethical Framework without delay', Press Release, March. https://www.unesco.org/en/articles/artificial-intelligence-unesco-calls-all-governments-implement-global-ethical-framework-without

Webb, C. 2020. The Impact of Artificial Intelligence on the Labor Market, mimeo. Stanford University.