White Paper

The Need for and Feasibility of an International AI Bill of Human Rights

By Professor Yuval Shany

Accelerator Fellowship Programme, Institute for Ethics in AI, University of Oxford

November, 2025

Introduction

The growing reliance of societies, economies and governments on AI systems in an ever-increasing range of fields and contexts has dramatic implications for the ability of individuals and groups of individuals to enjoy their basic human rights. An improved capacity to produce goods and deliver services in areas such as health, labour safety, transportation or education due to the increased productivity and accuracy of AI systems could potentially raise standards of living, as well as levels of human rights protection in these areas. At the same time, the disruptive effects of AI systems might put existing human rights under considerable pressure. For example, reliance by governments on the profiling features associated with certain AI systems they deploy could seriously harm the enjoyment of the right to non-discrimination and the right to privacy. What's more, the 'black box' features of AI systems and the resulting lack of transparency surrounding algorithmic decision-making processes invite consideration of whether our existing human rights norms should be revised so as to capture the new human needs and interests implicated by this new technology. A broad legal policy question confronts us in these and other comparable contexts: To what extent are existing human rights laws adequate to confront the challenges posed by the growing use of AI systems, and to what extent do they need revision in order to effectively respond to new challenges brought about by the increased deployment of AI systems?

To be sure, human rights laws have always contained reactive features, corresponding to new political, economic and technological realities. This is not surprising. Law is a social construct, and it continuously interacts with changing societal needs and expectations. For example, we have seen in recent years a growing acceptance at the international level of a new human right to a clean, healthy and sustainable environment in response to the deteriorating climate crisis.[1] Furthermore, international human rights law has developed over the years in ways that are intended to provide specific conduct-guidance to states and other relevant actors (including business entities), through the elaboration of more and more precise legal standards, adapting the general framework of human rights law to the needs of specific constituencies, like persons with disabilities who may find themselves in particular situations of vulnerability,[2] and in order to effectively tackle specific societal problems like corruption.[3]

It is against this background of AI disruption and a history of adaptation of human rights law, that this White Paper considers the question whether the latter should be adjusted to deal with the opportunities and risks created by the former. As this White Paper demonstrates, few if any effective processes of normative adjustment of international human rights law have occurred to date on the international plane. Still, a case can be made in favour of moving towards the elaboration of an “international AI bill of human rights” that would offer such an adjustment. The Paper maps, in this connection, prospects and challenges relating to the actual and potential human rights implications of AI systems – focusing on questions relating to access to AI systems, data protection, algorithmic bias and fairness, algorithmic transparency and explainability, manipulation, human decision and interaction, and accountability. It also considers shortcomings in the existing legal framework – the lack of AI-specific human rights norms that are fit-for-purpose and operate at an intermediate level of specificity, the need to avoid over-reliance on dynamic interpretation of existing international treaties by judicial and expert bodies, a concern about over-extending and over-loading current human rights norms (which were created with very different functional purposes in mind), and the inability of existing human rights law to fully capture the most important features that are unique to AI systems (e.g., ‘black boxes’ and lack of humanity) and to effectively govern the

[1] See e.g., UN General Assembly Resolution 76/300, The human right to a clean, healthy and sustainable environment, 28 July 2022, UN Doc. A/RES/76/300; Obligations of States in Respect of Climate Change, ICJ Advisory Opinion of 23 July 2025, at para. 393, https://www.icj-cij.org/sites/default/files/case-related/187/187-20250723-adv-01-00-en.pdf.

[2] Convention on the Rights of Persons with Disabilties, 12 Dec. 2006, UN Doc. A/RES/61/106 (hereinafter: CRPD).

[3] Human Rights Council Resolution 47/7, The negative impact of corruption on the enjoyment of human rights, 12 July 2021, UN Doc. A/HRC/RES/47/7 (2021).

operations of AI companies. The White Paper identifies a list of seven principles which should arguably be included in a future international AI bill of human rights and considers the way by which such a list can be developed – recommending the pursuit, at this stage, of informal (or “soft law”) norm creation processes.

Obviously, international human rights law already offers a large number of relevant standards which govern AI systems, such as the right to equality and the right to privacy. Furthermore, under the current international institutional framework, there exist a plethora of political, legal and professional policy-making venues for adapting, by way of interpretation, existing human rights norms to new technological developments, and for developing new international standards. Given their high degree of involvement in policy debates over contemporary problems, there were strong expectations that international human rights institutions would play a meaningful role in steering national, regional and international policies in the area of AI governance,[4] as well as in supporting efforts to protect individuals against violations of their rights by the use or failure to use AI systems. There were also growing expectations that international human rights bodies would develop effective ways to apply human rights standards and remedies to technology companies and find ways to hold them accountable for violating human rights.

To date, these expectations remain largely unmet. Although UN initiatives, such as the UN B-Tech,[5] the Secretary-General AI Advisory Body,[6] and the Working Group on Human Rights and Transnational Corporations,[7] like other multilateral[8] and

[4] See e.g., Kate Jones, AI Governance and Human Rights (Chatham House, 2023), https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights/04-principles-ai-governance-contribution-human-rights.

[5] https://www.ohchr.org/en/business-and-human-rights/b-tech-project.

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

[7] Report of the Working Group on the issue of human rights and transnational corporations and other business enterprises, Artificial intelligence procurement and deployment: ensuring alignment with the Guiding Principles on Business and Human Rights, 14 May 2025, UN Doc. HRC/59/53 (2025).

[8] See e.g., Hiroshima Process International Code of Conduct for Advanced AI Systems, 30 Oct. 2023, https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems.

multistakeholder initiatives,[9] have been taking increased interest in AI and human rights, and while several UN human rights officials have expressed themselves on AI governance issues,[10] no detailed, fit-for-purpose, comprehensive and broadly-supported human rights roadmap has emerged to date for dealing with the key challenges of applying and adapting international human rights law to new challenges linked to AI systems. Most international documents promulgated so far in this policy space tend to address human rights questions at a high level of abstraction – typically at the level of general principles – or at a very granular and technical regulatory level. The latter regulation tends not to be articulated in the language of universal human rights standards.

Of course, the act of transforming policy recommendations in the field of AI governance into concrete international human rights standards is complicated and should be approached with great care. **First**, certain dimensions of the applicable policies would need to be expressed in the language of international human rights law, in ways that maintain the latter system's need for coherence, meet its traditional definitions, principles and structural categories (or "grammar"),[11] and be compatible with key normative assumptions underlying international human rights law. We should recall, in this connection, that many AI governance policy recommendations are grounded in domestic or regional legal policy instruments, and not in universal human rights law. Ultimately, there is a need to assess whether and how existing human rights should be re-interpreted so as to capture rights-protecting AI policies, or whether new specific rights need to be identified and developed in order to effectively anchor the relevant human rights-centred AI policies. One may note however, in this regard, the aversion of many human rights actors to developing new rights due to fears of "rights inflation", and

[9] See e.g., The Global Partnership on Artificial Intelligence (GPAI), Responsible AI Working Group Report, Dec. 2023, https://www.gpai.ai/projects/responsible-ai/Responsible%20AI%20WG%20Report%202023.pdf.

[10] See e.g., Report of the Special Rapporteur on extreme poverty and human rights, Extreme poverty and human rights, UN Doc. A/74/493 (2019), p. 16, https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=A/HRC/49/52&Lang=E.

[11] See e.g., Bonavero Institute of Human Rights, Addressing the digital realm through the grammar of human rights law (2020-2025), https://www.law.ox.ac.uk/addressing-the-digital-realm-through-the-grammar/addressing-digital-realm-through-grammar-human.

concerns about the legal and political risks associated with acknowledging existing normative gaps.

Second, there might be a need for developing modalities for formulating and applying international human rights law standards relating to the use of AI systems in ways that deviate from traditional human rights paradigms. This includes the application of international human rights law to major tech companies that design, develop, manufacture, operate and disseminate AI systems, and have significant influence on the ways in which these systems impact the enjoyment of human rights by individuals. Application of human rights to private actors is, however, a major challenge for international human rights law which has been historically created to primarily regulate state activity. It remains to be seen how much progress can be made within the current (arguably, under-effective) international human rights law framework for business and human rights,[12] in order to enhance the normative pull exerted by international human rights law norms on private companies in the AI governance context. Furthermore, the dispersed and non-transparent nature of the many of the adverse impacts generated by the use or failure to use AI systems renders the individual victim-centred process of enforcement of international human rights law practiced by many human rights bodies largely ineffective.

One possible response to these challenges is the development of deliberate initiatives for clarifying and developing comprehensive international human rights standards for the use of AI systems, and finding ways to introduce them into the relevant human rights policies and practices of leading tech companies working in the field. Such human rights policies, and their actual application to specific products and services, could be regularly monitored by internal or external mechanisms of review (such as those created pursuant to the EU Corporate Sustainability Due Diligence Directive).[13] In the course of such reviews, concerns and complaints regarding the implementation of specific human rights standards, and even the adequacy of the relevant human rights

[12] See UN Guiding Principle on Business and Human Rights (2011)(hereinafter: UNGPs), https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

[13] See Directive (EU) 2024/1760 of the European Parliament and of the Council of 13 June 2024 on Corporate Sustainability Due Diligence, 5 July 2024, OJ L, 2024/1760.

policies, could be raised. Note however, that some of the major companies in the field of AI have adopted to date only very general human rights policies that lack detailed language on which specific human rights norms apply to their AI products and services, and how such norms apply.[14] Nor do these companies tend to provide for independent complaints mechanisms, which could robustly enforce human rights standards in appropriate cases. Increasing normative clarity and specificity in this policy space through an international AI bill of human rights might arguably help AI companies in assuming a greater part of the burden associated with application of human rights to use cases involving AI systems. It may also facilitate the articulation by a variety of stakeholders and social actors of demands for policy change in the direction of greater human rights accountability by AI companies.

In undertook the research conducted to prepare this White Paper during my tenure as an inaugural fellow at the Oxford Ethics in AI Institute's Accelerator Fellowship (2024-2025). I was greatly assisted in my research by four high level consultations coordinated by the Accelerator Fellowship Programme in cooperation with four other research centres – the Bonavero Institute of Human Rights at Oxford, the Geneva Academy of International Humanitarian Law and Human Rights, the Harvard Human Rights Program and the Pretoria Centre for Human Rights. While the responsibility for the conclusions and recommendations offered below is mine and mine alone, my work greatly benefited from words of advice, criticisms and ideas discussed during the consultations (which were all conducted in accordance with the Chatham House Rules).

The White Paper maps opportunities and challenges relating to the actual and potential human rights implications of the use of AI systems, and considers shortcomings in the existing legal framework. It identifies a list of seven principles which should arguably be included in a future international AI bill of human rights, and considers the means by which such a list can emerge, recommending the pursuit, at this stage, of informal (or

[14] See e.g., https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf; https://ai.google/responsibility/principles/; https://www.hpe.com/emea_europe/en/solutions/artificial-intelligence/ethics.html.

“soft law”) norm creation processes. Part One briefly discusses several areas of interaction between uses of AI systems and human rights, which may invite some adaptation or development of existing norms. Part Two analyses a number of recent standard-setting instruments adopted at the national, regional and global level, which address, directly or indirectly, the protection of individual rights in the face of uses of AI systems. Part Three evaluates the quality of human rights protections afforded by these existing standard-setting instruments and explains, in light of the consultations held as part of the research programme, why an international AI bill of human rights could fill in existing normative gaps and promote a more effective system of human rights protections for the age of AI. Part Four considers practical avenues for advancing the articulation and adoption of an international AI bill of human rights and elaborates the potential contents of such an instrument. Part Five concludes.

Acknowledgement

I wish to thank the staff and members of the Oxford Ethics in AI Institute – Accelerator Fellowship Programme for their excellent support for the research, and to participants, institutional partners and co-organizers of the consultations undertaken as part of the research programme. The research also benefited from the support received from members of my ERC research group, which is funded by ERC Grant No. 101054745: the Three Generations of Digital Human Rights (DigitalHRGeneration3), https://3gdr.huji.ac.il/.

I. Human rights challenges posed by AI systems

Scope, speed and scale

It is hardly surprising that the disruptive nature of AI technology affects many private and public interactions involving human beings, implicating many of their human rights. Some of these effects do not pose a serious conceptual challenge. Still, they raise many practical difficulties which merit attention and treatment. To the extent that AI systems replace human beings, many issues similar to those raised in connection with human acts or omissions may now arise with regard to acts or omissions mediated through the substituting AI systems. For example, free speech restrictions can occur either through human-operated content moderation of social media posts or through AI-operated content-moderation. In the same vein, privacy infringements can occur through human-operated wiretapping devices or through AI-run digital surveillance. The distinction between these two types of rights-infringements often involves a difference in scope, speed and scale: Human content moderation teams move slower than AI systems when removing offensive speech from social media, and AI technology can support far more intrusive and extensive use of surveillance systems by governments (including, for instance, facial recognition and emotion recognition systems), as well as greater capacity to engage in decryption, de-anonymization of data and tracing back its origins.

Furthermore, the use of AI systems may generate, from a human rights perspective, new problems, which might exceed the scope of the effective protection schemes currently afforded by human rights law. To illustrate, the dependency of AI systems on ever-expanding training data and post-training data introduces an unprecedented demand for data collection that, in and of itself, generates a strong incentive for privacy breaches. As explained below, the use of AI systems also entails problems of transparency and accountability. As a result, the replacement of human beings by AI systems could render the associated privacy-related infringement more serious and harder to tackle, even if AI systems undertake essentially the same functions which humans undertook before. We might need, in such cases, stronger reasons to justify the use of AI systems for tasks we can justify when undertaken by humans.

Safety and accountability

Another way by which resort to AI systems may result in more serious human rights harms than those caused by humans involves accidents, malfunction or foul play. To be sure, faulty or inaccurate use of important health care or financial management systems can cause significant harm to impacted individuals regardless of whether these systems are operated by humans or by AI systems, and both sets of harms may undercut those individuals' ability to enjoy a variety of human rights, such as the right to life or health, the right to peaceful enjoyment of property and the right to adequate standard of living. Here too, however, there are particular concerns relating to the deployment of AI systems. These include problems relating to the technological reliability and the particular vulnerability of AI systems to hacking and other forms of sabotage, the risk that AI systems – lacking in common sense[15] and having an idiosyncratic 'worldview'[16] – would commit catastrophic mistakes that no human is ever expected to make (such as operating on the wrong body organ or giving away for free someone's entire assets), and problems of holding anyone accountable for harms caused by AI systems. Indeed, the problem of accountability for the use of AI systems raises a particularly difficult challenge from a human rights perspective. The difficulty in identifying a specific human being responsible for acts and omissions undertaken by an algorithm – especially in cases where AI systems develop across a long value chain and operate in unpredictable ways – might erode the right to an effective remedy, which is a procedural "secondary" human right to have "primary" human violations (that is, violations of substantive rights, such as the right to life, health, privacy or equality) properly addressed by states and other duty holders.[17]

[15] See e.g., Martin W. Bauer and Bernard Schiele, When Artificial Intelligence Meets Common Sense, Frictions will Arise, in *AI and Common Sense: Ambitions and Frictions* (Martin W. Bauer and Bernard Schiele eds., 2024) 3.

[16] See e.g., Alberto Romele, A. The datafication of the worldview, 38 *AI & Soc* (2023) 2197.

[17] For the distinction between rules of primary and secondary order, see HLA Hart, *The Concept of Law* 3rd ed., 2012; originally published in 1961) Ch. V.

Better quality of services

The flip side of the differences between the quality of services performed by human beings and those assisted or fully undertaken by AI systems may be that impacted individuals may have good reasons to insist that certain services offered to them would involve AI systems that meet higher-than-human quality standards. These could be particularly relevant with regard to AI systems used in the health sector, but also in connection with systems deployed in education programs, systems ensuring work place safety, and systems providing services for the disabled and prompt access to public benefits. Demands for enjoying the dividends associated with the use of AI systems could be sometimes articulated in the language of rights: For instance, it may be claimed that resort to AI systems is dictated by the right to "the enjoyment of the highest attainable standard of physical and mental health",[18] or the right of persons with disability to "have access to a range of in-home, residential and other community support services, including personal assistance necessary to support living and inclusion in the community".[19] As further discussed below, such demands can also be linked to the (relatively obscure) right to "enjoy the benefits of scientific progress and its applications".[20] It may be also claimed that the use of AI systems constitutes part of the effective protection of certain human rights and part of the effective remedy that duty-holders should afford victims of human rights violations (e.g., quick access to compensation schemes).

Discrimination

One specific area where the use of AI systems introduces a new set of challenges is the field of anti-discrimination laws and policies, where AI systems have been alleged to display algorithmic bias and lack of fairness,[21] often involving not only the perpetuation of pre-existing inequities in society reflected in the training or post-training data used for the development of AI systems, but also in new forms of inequality caused by the

[18] International Covenant on Economic, Social and Cultural Rights, 16 Dec. 1966, art. 12, 993 UNTS 3 (hereinafter: ICESCR).

[19] CRPD, art. 19(b).

[20] ICESCR, art. 15(1)(b).

[21] See e.g., Nima Kordzadeh and Maryam Ghasemaghaei, Algorithmic Bias: Review, Synthesis, and Future Research, 31 *European Journal of Information Systems* (2021) 388.

manner in which training and post-training data is collected – overrepresenting certain population groups and underrepresenting other population groups – and by the way in which data is actually used by algorithms – prioritizing certain quantitative factors over other quantitative and qualitative factors in AI-enabled decision making systems. At a more general level, the treatment of individuals on the basis of their group affiliation – which correlation-based AI systems are trained to identify and issue predictions in relation of – often stands in tension with notions of human dignity that invite treating individuals respectfully as unique and full-fledged human beings, regardless of their group affiliation and their wish to 'un-belong' to any particular group and avoid being saddled with group stereotypes. Moreover, problems of lack of transparency and the use of many indirect proxies for 'suspect' or prohibited group classifications[22] render instances of algorithmic discrimination harder to trace and to hold those responsible for it to account than when discrimination occurs in non-AI contexts.

Black box

As discussed below, there may be certain contexts in which the contents of international human rights law might need to change in order to accommodate new features associated with the use of AI systems, which have no immediate parallel in a world without AI. One of these contexts involves problems arising out of the 'black box' attributes of AI systems. Given the non-intelligibility of machine algorithms to most users of AI systems, and the unpredictability – even to system developers – of outcomes generated through 'deep learning' AI systems operating on the basis of unsupervised machine learning (often involving multiple layers of neural networks),[23] algorithmic decisions and the reasons underlying them frequently remain opaque.

Such lack of transparency has significant human rights implications. Among other things, it greatly complicates the ability of human rights victims to identify the violations

[22] For a discussion of 'suspect classification', see Marcy Strauss, Reevaluating Suspect Classifications, 35 *Seattle University Law Review* (2011) 135.

[23] See e.g., Juha Karhunen, Tapani Raiko and Kyung Hyun Cho, Unsupervised deep learning: A short review, in *Advances in Independent Component Analysis and Learning Machines* (Ella Bingham, Samuel Kaski, Jorma Laaksonen and Jouko Lampinen, eds., 2015) 125.

they have been made subject to (e.g., AI biases) and to hold perpetrators (e.g., developers and deployers) accountable. Furthermore, the failure to anticipate machine outcomes and to understand the reasons underlying them, paints such decisions as arbitrary and unfair in nature in the eyes of those impacted by them. This might run afoul of specific duties to provide reasons as part of procedural fairness and substantive justice guarantees. In international human rights law, such normative expectations tend to be reflected in the language of certain treaty provisions (which include, for example, prohibitions on arbitrary interference with privacy[24] and arbitrarily detention,[25] and a duty on courts to provide legal reasoning for their judgments).[26] It could be argued, in this connection, that the movement away from human reasoning creates new problems of arbitrariness, unfairness and injustice, which justifies the development of more robust transparency safeguards to offset the process of disempowerment of humans impacted by decisions rendered through the use of AI systems.

What's more, the combined effect of the 'black box' features of sophisticated AI systems and the increasingly central role such systems play in the lives of individuals and groups of individuals, render the way in which the world runs less and less comprehensible to many individuals, whose ability to exercise meaningful agency in such a world is eroding. This epistemic crisis,[27] which often manifests itself also as a democratic crisis due to the knock-on effects of lack of knowledge on the ability to rationally formulate political preferences, may also justify a human rights response designed to enable individuals to better navigate the societal and informational context in which their rights are now supposed to be exercised.

[24] International Covenant on Civil and Political Rights, 16 Dec. 1966, art. 17, 999 UNTS 171 (hereinafter: ICCPR).
[25] ICCPR, art. 9.
[26] Human Rights Committee, General Comment No. 31: Article 14: Right to equality before courts and tribunals and to a fair trial, para. 29, UN Doc. CCPR/C/GC/32 (2007).
[27] See e.g., Aaron Hyzen, Hilde Van den Bulck, Manuel Puppis, Michelle Kulig and Steve Paulussen, Epistemic welfare and algorithmic recommender systems: overcoming the epistemic crisis in the digitalized public sphere, *Communication Theory* (2025), https://academic.oup.com/ct/advance-article/doi/10.1093/ct/qtaf018/8240891.

Manipulation and impersonation

The decreasing ability of individuals to make sense of a world that is mediated through AI systems and to distinguish between authentic contents and the inauthentic contents such systems generate (e.g., information v. disinformation or misinformation; authentic videos v. 'deep fakes'), creates ripe conditions for an extensive use of AI systems for manipulative purposes. The risk of illegitimate manipulation attaches to deceitful misrepresentations of reality, but also to the use of AI systems in order to apply subtle nudges and subliminal influences that push individuals to act or refrain from acting in certain ways, in a manner that erodes or circumvents rational thinking and conscious deliberation. This could, and often does result in individuals conducting themselves in ways that are contrary to their best interests.[28] In other words, the use of AI system to disrupt data integrity and mental integrity constitutes a threat to individual agency and autonomy, and could threaten a number of human rights, including freedom of thought, freedom of opinion, and the right to seek and receive information, as well as other human rights that could be adversely affected by bad choices facilitated by manipulative uses of AI systems, such as the right to take part in the conduct of public affairs.

One particular risk of manipulation involves the risk of impersonation – that is, interaction of humans with AI systems without being aware of the artificial nature of the latter. Impersonation may create a heightened risk of illegitimate acts of deceit – through propagation of misinformation/disinformation – and illegitimate nudges. It also puts at risk relationships of trust between interlocutors which could compromise the privacy rights of individuals, as well as other human rights that depend on trust in fields like health, education, welfare and criminal justice.

Human decision and interaction

Finally, the aforementioned potential flaws of AI systems – inaccuracy, non-transparency, non-accountability, arbitrariness and reliance on discriminatory profiling – could render it a sub-optimal decision-maker in the eyes of many individuals. Add to

[28] For a discussion, see Cass Sunstein, *Manipulation: What it is, Why it's bad, What to do about it?* (2025) 164-195 (hereinafter: Sunstein, *Manipulation*).

that, the inhumanity of AI systems – their inability to experience emotions, to feel empathy, to authentically relate to humans on a personal level and to regard individuals interacting with them as full-fledged human beings (as opposed to a mere collection of data points). This could give rise to demands by individuals to opt out from decision-making processes undertaken by AI systems or be able to challenge the decisions they generated before human beings. The demand to opt out from the decisional authority of AI systems has obvious human rights dimensions when the decisions in question impact the enjoyment of human rights (for instance, eligibility for welfare benefits or education, or adjudication of legal responsibility). Still, it may be suggested that the demand not to be subject to automated decisions touches upon basic human dignity considerations – the right to be treated with respect as a human person – in ways that could justify the development of human rights norms to address the very existence of automated decisional authority.

In the same vein, it can be claimed that certain sensitive interactions implicating existing human rights, such as interactions around medical or psychological treatment, educational activities, old age or disability care and child rearing, should not be offered exclusively through AI systems when the affected individuals oppose it. The need to maintain the possibility for human-to-human interaction in key areas of life could be another area to which human rights norms reflective of the need to protect human dignity would eventually extend.

While not presuming to be exhaustive in nature, this overview survey of potential human rights problems associated with the increased use of AI systems suggests that there is a need to consider whether existing human rights norms ought to be adjusted or supplemented in order to deal with these new or enhanced human rights challenges. The next part of the White Paper (Part Two) examines recent standard-setting instruments that purport to adjust or supplement existing human rights standards through the development of specific legal norms in this area. A subsequent part (Part Three) evaluates the capacity of these instrument to afford protections equivalent to those found in international human rights instruments.

II. Recent standard-setting initiatives

A large number of standard-setting instruments were concluded in recent years, with a view to addressing gaps and inadequacies in existing legal frameworks that are supposed to protect individual needs and interests adversely affected by new AI technologies. Some of them have even introduced explicitly or implicitly new legal rights relevant to the use of AI systems. These include EU regulations on data privacy (the 2016 General Data Protection Regulation or GDPR)[29] and AI (the 2024 AI Act),[30] the Council of Europe Convention on AI (the 2024 Framework AI Convention),[31] and the now-abandoned 2022 White House Blueprint for an AI Bill of Rights.[32] The list of relevant standard-setting instruments is long, and keeps getting longer. It now includes global instruments (such as the UN Global Digital Compact from 2024),[33] new regional initiatives (such as the African Union's 2024 Continental AI Strategy)[34] and, increasingly, national legislation (such as the 2025 South Korean Basic AI Act[35] and the 2025 Italian AI Law).[36] Interestingly, for our purposes, these different instruments tend to coalesce around certain legal protections that are common to most of them, if not all

[29] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of such Data, 4 May 2016, OJ L 119, at 1 (hereinafer: GDPR).

[30] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, 12 July 2024, OJ L 2024/1689, http://data.europa.eu/eli/reg/2024/1689/oj (hereinafter: AI Act).

[31] The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 5 Sept. 2024, art. 9, ETS 225 (hereinafter: CoE Framework AI Convention). The Framework AI Convention has not entered into force at the time of writing.

[32] The White House Office of Science and Technology Policy (OSTP), Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People (October, 2022), https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/ (hereinafter: White House Blueprint).

[33] UN General Assembly, Resolution 79/1, Pact for the Future, Annex I: Global Digital Compact, 22 Sept. 2024, UN Doc. A/RES/79/1 (2024)(hereinafter: Global Digital Compact).

[34] AU, Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity (July 2024), https://au.int/sites/default/files/documents/44004-doc-EN-_Continental_AI_Strategy_July_2024.pdf (hereinafter: AU Continental AI Strategy).

[35] Basic Act on the Development of Artificial Intelligence and Establishment of Trust, Law No. 20676 (Republic of Korea), 21 Jan. 2025.

[36] Provisions and Powers Delegated to the Government concerning Artificial Intelligence (Italy), 17 Sept. 2025, https://www.senato.it/service/PDF/PDFServer/BGT/01462298.pdf?utm_source=chatgpt.com (hereinafter: Italian AI Law). I relied on an unofficial translation of the law.

of them. These common normative arrangements may suggest, as I explain below, a growing international support for the emergence of certain AI human rights.

Access to AI systems

One common theme found in some standard-setting instruments involves access to digital technologies. For example, the Global Digital Compact proclaims on behalf of the entire UN membership that:

> Accessible and affordable data and digital technologies and services are essential to enable every person to participate fully in the digital world. Our cooperation will promote digital accessibility for all and support linguistic and cultural diversity in the digital space;[37]

In the same spirit, the UN member states made the following commitment regarding access to AI systems:

> We will leverage existing United Nations and multi-stakeholder mechanisms to support artificial intelligence capacity-building to bridge artificial intelligence divides, facilitate access to artificial intelligence applications and build capacity in high-performance computing and related skills in developing countries.[38]

Language on access to AI systems can also be found in the South Korean Basic AI Act ("Ensuring that products and services powered by AI technology are freely and conveniently accessible by everyone"),[39] in the Italian AI Law[40] and in the concluding statements of some international AI summits (for example, the 2025 Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet identified as a priority "[p]romoting AI accessibility to reduce digital divides").[41] Such language, calling

[37] Global Digital Compact, para. 8(g).

[38] Global Digital Compact, para. 61.

[39] South Korean Basic AI Act, art. 27(2). See also Italian AI Law, art. 7-14 (promoting or allowing for the use of AI in a variety of sectors, including economy, education, health, disability, public administration, labor, judicial activity, criminal investigation).

[40] Italian AI Law, art. 4(6)(duty to ensure accessibility by persons with disability to AI systems).

[41] AI Action Summit, Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet, 11 Feb. 2025, https://onu.delegfrance.org/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and.

on states to ensure access to AI systems, mirrors the language used in digital and Internet rights charters and declarations adopted by some countries and by the EU,[42] which reaffirmed a right to access digital technology (see e.g., declarations or charters adopted by Brazil,[43] Italy,[44] Spain[45] and Portugal).[46]

It is notable in this regard that the 2022 EU Declaration alludes in this regard to access to a "trustworthy, diverse and multilingual digital environment",[47] and that the sentiment that access has to meet certain quality requirements is also echoed in the Global Digital Compact ("[f]oster an inclusive, open, safe and secure digital space that respects, protects and promote human rights").[48] This suggests that the right to access digital services, including those facilitated by AI systems, has to be developed in particular ways that respect the needs and interests of users, including safety considerations.

AI Safety

Indeed, notions of safety, reliability and resilience, which correspond to the aforementioned concerns about catastrophic mistakes that the use of AI systems could cause, as well as to additional concerns about hacking, malfunctions and other unintended consequences of AI 'frontier' technology, do appear in a number of standard-setting instruments which articulate or allude to individual rights relating to the use of AI systems. Such notions are sometimes presented as independent normative requirements attached to the development and deployment of AI systems. The AI Act is designed, for example, "to promote the uptake of human centric and trustworthy artificial

[42] European Declaration on Digital Rights and Principles for the Digital Decade, 15 Dec. 2022, OJ C 23/1 (2023) (hereinafter: EU Delcaration).

[43] Marco Civil Law of the Internet, Law No. 12.965 (Brazil), 23 April 2014, https://www.cgi.br/pagina/marco-civil-law-of-the-internet-in-brazil/180.

[44] Declaration of Internet Rights (Italy), 28 July 2015, https://www.camera.it/application/xmanager/projects/leg17/commissione_internet/testo_definitivo_inglese.pdf.

[45] Charter of Digital Rights (Spain), 14 July 2021, https://portal.mineco.gob.es/RecursosArticulo/mineco/ministerio/participacion_publica/audiencia/ficheros/Charter%20of%20Digital%20Rights.pdf (hereinafter: Spanish Charter).

[46] Charter on Human Rights in the Digital Age, Law No. 27/2021 (Portugal), 17 May 2021, https://diariodarepublica.pt/dr/detalhe/lei/27-2021-163442504 (hereinafter: Portuguese Charter).

[47] EU Declaration, para. 12.

[48] Global Digital Compact, para. 7(3).

intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union… to protect against the harmful effects of AI systems in the Union, and to support innovation”.[49] The White House Blueprint is even more explicit in proclaiming a right to AI safety: “You should be protected from unsafe or ineffective systems”.[50] And the Framework AI Convention provides that “[e]ach Party shall take, as appropriate, measures to promote the reliability of artificial intelligence systems and trust in their outputs, which could include requirements related to adequate quality and security throughout the lifecycle of artificial intelligence systems”.[51]

The Global Digital Compact calls, in this regard, on “standards development organizations to collaborate to promote the development and adoption of interoperable artificial intelligence standards that uphold safety, reliability, sustainability and human rights”.[52] In the same vein, during the 2023 AI safety summit, participating states agreed on the following action items:

- identifying AI safety risks of shared concern, building a shared scientific and evidence-based understanding of these risks, and sustaining that understanding as capabilities continue to increase, in the context of a wider global approach to understanding the impact of AI in our societies.
- building respective risk-based policies across our countries to ensure safety in light of such risks, collaborating as appropriate while recognising our approaches may differ based on national circumstances and applicable legal frameworks. This includes, alongside increased transparency by private actors developing frontier AI capabilities, appropriate evaluation metrics, tools for

[49] AI Act, rectial para. 1.
[50] White House Blueprint, p. 15.
[51] CoE AI Framework, art. 12.
[52] Global Digital Compact, para. 58.

> safety testing, and developing relevant public sector capability and scientific research.[53]

Finally, the South Korean Basic AI Act calls for developing "ethics principles" for "[s]afety and reliability to prevent harm to human life and physical and mental health during the development and use of AI".[54]

Data protection and data privacy

AI systems introduce new significant risks for existing privacy and data protection regimes. Such risks emanate not only from the ability to harness AI systems for massive data collection and data analysis operations with significant privacy implications (exacerbated by the capacity of AI systems to engage in decryption and de-anonymization), but also because of the insatiable demand of the AI systems themselves for vast quantities of data. At the same time, the lack of transparency around the operation of AI systems renders it difficult to ascertain by way of 'reverse-engineering' what data was actually collected and how such data might have been used. This greatly complicates the ability of individuals to enforce their privacy and data protection rights. It is therefore not surprising that standard-setting instruments in the field of AI emphasize the need for developing and using AI systems in accordance with effective data protection standards.

The need for effective data protection and privacy standards has been articulated in the Global Digital Compact:

> We recognize that responsible and interoperable data governance is essential to advance development objectives, protect human rights, foster innovation and promote economic growth. The increasing collection, sharing and processing of

[53] The Bletchley Declaration by Countries attending the AI Safety Summit, 2 November 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

[54] South Korean Basic AI Act, art. 27(1).

> data, including in artificial intelligence systems, may amplify risks in the absence of effective personal data protection and privacy norms.[55]

The centrality of privacy standards for activities of AI systems has also been acknowledged in the Council of Europe's Framework AI Convention, which provides as follows:

> Each Party shall adopt or maintain measures to ensure that, with regard to activities within the lifecycle of artificial intelligence systems: (a) privacy rights of individuals and their personal data are protected, including through applicable domestic and international laws, standards and frameworks; and (b) effective guarantees and safeguards have been put in place for individuals, in accordance with applicable In and international legal obligations.[56]

In the same vein, the EU AI Act mandates that "AI systems are developed and used in accordance with privacy and data protection rule"[57] (including the GDPR), and the AU's Continental AI Strategy calls on state to increase their cooperation around data protection and data governance.[58]

Adopting an even more explicit stance on the matter, the White House Blueprint recognizes a right to data privacy in connection with the use of AI systems. It suggests that "[y]ou should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used".[59] Similarly, the 2025 Italian AI Law provides for data protection safeguards relating to AI systems, including "the principles of lawfulness, fairness, transparency, accuracy, and limitation of purpose".[60]

[55] Global Digital Compact, para. 37.
[56] CoE Framework AI Convention, art. 11.
[57] AI Act, recital para. 27.
[58] AU Continental AI Strategy, p. 46.
[59] White House Blueprint, p. 30.
[60] Italian AI Law, art. 6(1).

Non-discrimination

The use of AI systems carries serious discrimination risks stemming from the non-transparent nature of such systems, their reliance on problematic datasets – which may reflect or even exacerbate pre-existing practices of discrimination – the propensity of such systems to use group characteristics as predictive features and their bias towards reliance on quantitative factors in decision-making in a manner that often benefits members of certain groups at the expense of other groups. As a result, standard-setting instruments dealing with AI systems often contain language on protecting the right to equality and curbing discrimination.

For example, the Global Digital Compact calls to protect users of digital technology from "violations, abuses and all forms of discrimination",[61] the EU Declaration commits to "ensuring that algorithmic systems are based on adequate datasets to avoid discrimination"[62] and the AU Continental Strategy emphasizes the "need for legal protection against algorithmic bias and discrimination". The Council of Europe Framework AI Convention explicitly provides the following legal requirements:

1. Each Party shall adopt or maintain measures with a view to ensuring that activities within the lifecycle of artificial intelligence systems respect equality, including gender equality, and the prohibition of discrimination, as provided under applicable international and domestic law.
2. Each Party undertakes to adopt or maintain measures aimed at overcoming inequalities to achieve fair, just and equitable outcomes, in line with its applicable domestic and international human rights obligations, in relation to activities within the lifecycle of artificial intelligence systems.[63]

The White House Blueprint also states that "[y]ou should not face discrimination by algorithms and systems should be used and designed in an equitable way",[64] and the AI

[61] Global Digital Compact, para. 22.

[62] EU Declaration, para. 9(c).

[63] CoE Framework AI Convention, art. 10.

[64] White House Blueprint, p. 23. See also Italian AI Law, art. 4(2)(enumerating non-discrimination among applicable general principles) and art. 9, 10, 12, 14 (affirming the application of non-discrimination to uses of AI in the health, disability, labor and surveillance sectors).

Act requires states to subject high-risk AI systems (e.g., systems that deploy biometrics or operate in sensitive sectors like law enforcement, education, immigration and essential public services) to "examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations".[65] One may also note in this connection the concern about profiling of individuals found in both the AI Act[66] and the GDPR,[67] and the call for "data justice" found in the African Union's 2022 Data Policy Framework.[68]

Algorithmic transparency

As indicated above, the 'black box' features of AI systems render some AI-enabled decisions arbitrary and unfair in the eyes of those individuals affected by them. Such lack of transparency also complicates the ability of victims to identify and denounce violations of their human rights, including the right to privacy and non-discrimination. It also increases risks of AI manipulation.

Concerns of this kind find an expression in the White House Blueprint that affirms a right to algorithmic transparency: "You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you".[69] In the same vein, the Council of Europe Framework AI Convention provides that "[e]ach Party shall adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place in respect of activities within the lifecycle of artificial intelligence systems, including with regard to the identification of content generated by artificial intelligence systems".[70]

These standards sit well with those found in the EU Declaration, where EU members have committed to "ensuring an adequate level of transparency about the use of

[65] AI Act, art. 19(2)(f).
[66] AI Act, recital para. 42.
[67] GDPR, art. 22.
[68] https://au.int/sites/default/files/documents/42078-doc-AU-DATA-POLICY-FRAMEWORK-ENG1.pdf.
[69] White House Blueprint, p. 40.
[70] CoE Framework AI Convention, art. 8.

algorithms and artificial intelligence, and that people are empowered to use them and are informed when interacting with them".[71] A roughly similar approach has been taken in the South Korean Basic AI Act ("AI business operators providing products or services using high-impact AI or generative artificial intelligence (GenAI) shall notify users in advance that the product or service is AI-based"),[72] the Italian AI law ("1. Citizens must be informed in a clear and accessible way about the use of artificial intelligence by public administrations and companies. 2. Labels, notices, or other forms of transparency must indicate when a citizen interacts with an artificial intelligence system"),[73] and the GDPR.[74]

The AI Act further requires that transparency be baked into certain AI systems: "High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately". As for the scope of the applicable transparency obligations, the AI Act provides that transparency be guaranteed in regard to the artificially generated or manipulated nature of the contents generated (including 'deep fakes' and 'fake news') and the use of emotion recognition systems.[75] With regard to AI-enabled decisions involving high-risk systems, the AI Act imposes a duty of explanation "of the role of the AI system in the decision-making procedure and the main elements of the

[71] EU Declaration, para. 9(b). See also Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act), 27 Oct. 2022, art. 27(1), *OJ L 277/1* (hereinafter: DSA) ("Providers of online platforms that use recommender systems shall set out in their terms and conditions, in plain and intelligible language, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters").

[72] Korean Basic AI Act, art. 31(1).

[73] Italian AI Law, art. 15(1)-(2)(unofficial translation).

[74] GDPR, art. 12(1)("The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means. When requested by the data subject, the information may be provided orally, provided that the identity of the data subject is proven by other means"). See also African Union Convention on Cyber Security and Personal Data Protection (Malabo Convention), 27 June 2014, art. 13, principle 5, https://ccdcoe.org/uploads/2018/11/AU-270614-CSConvention.pdf ("The principle of transparency requires mandatory disclosure of information on personal data by the data controller"); Measures for the Administration of Internet Information Services on the Labeling of Synthetic Content Generated by Artificial Intelligence (PRC), 7 March 2025, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

[75] AI Act, art. 50.

decision taken".[76] This mirrors the aforementioned approach of the White House Blueprint on the right to explanation.[77] Finally, the AI Act singles out particular concerns about the risk of impersonation due to lack of transparency about the artificial features of the system used.[78]

Note that demands for greater transparency and accountability often go hand-in-hand in the relevant standard-setting instruments. The Global Digital Compact calls, for example, to "[p]romote transparency, accountability and robust human oversight of artificial intelligence systems in compliance with international law",[79] and the White House Blueprint lays out, under its "notice and explanation" section, expectations for provision of information on "the individual or organization responsible for the system."[80] By contrast, the Framework AI Convention provides separately for transparency (and oversight) and for accountability and responsibility. [81]

Manipulation

One specific concern expressed by standard-setting instruments seeking to protect individuals from certain abusive uses of AI systems pertains to the risk of manipulation. AI systems may be particularly effective in exploiting vulnerabilities, constraining free choices, providing misinformation or disinformation and employing 'dark patterns'.[82] The general risk of manipulation in the context of harms to informational integrity in the digital age has been expressed in the Global Digital Compact ("We will strengthen international cooperation to address the challenge of misinformation and disinformation and hate speech online and mitigate the risks of information manipulation in a manner consistent with international law").[83] In the same vein, the EU Declaration calls on the

[76] AI Act, art. 86.

[77] White House Blueprint, p. 40.

[78] AI Act, recital para. 132;

[79] Global Digital Compact, para. 55(d).

[80] White House Blueprint, p. 40.

[81] CoE Framework AI Convention, art. 9 ("Each Party shall adopt or maintain measures to ensure accountability and responsibility for adverse impacts on human rights, democracy and the rule of law resulting from activities within the lifecycle of artificial intelligence systems").

[82] Dark patterns are deceptive techniques used to manipulate choices. European Parliamentary Research Service, Regulating dark patterns in the EU: Towards digital fairness (2025), https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/767191/EPRS_ATA(2025)767191_EN.pdf.

[83] Global Digital Compact, para. 33. See also Portuguese Charter, art. 11 (right to protection against disinformation).

member states to create digital and cyber environments that protect people from manipulation,[84] the AU Continental AI Strategy expresses concern about the need to protect informational integrity,[85] and the Spanish Digital rights charter contains protections against identity manipulation and manipulating the will of minors.[86]

The AI Act explicitly prohibits the use of AI systems engaged in purposeful and harmful manipulative techniques:

> The following AI practices shall be prohibited: (a) the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm;[87]

Chinese regulations on 'deep fakes' and algorithmic recommendation systems appear to similarly protect users against being manipulated by false images or tendentious recommendations,[88] and the commentary to the White House Blueprint stipulates that "[u]ser experience design choices that intentionally obfuscate or manipulate user choice (i.e., "dark patterns") should not be used".[89] Implicit support for the same approach can also be surmised from the protection of human dignity and individual autonomy found in the Framework AI Convention.[90]

[84] EU Declaration, para. 15(d), 16(b).

[85] AU Continental AI Strategy, pp. 49-51.

[86] Spanish Charter, art. II (2)-(3), X(3).

[87] AI Act, art. 5(1).

[88] Provisions on the Administration of Deep Synthesis Internet Information Services, 12 Dec. 2022 (China), https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm; Regulations on the Management of Algorithm Recommendation of Internet Information Services, 31 Dec. 2021(China), https://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm

[89] White House Blueprint, p. 34. See also DSA, art. 25(1)("Providers of online platforms shall not design, organise or operate their online interfaces in a way that deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions").

[90] CoE Framework AI Convention, art. 7.

Automated decisions

One specific concern which almost all standard-setting instruments share relates to the delegation of important decisions from humans to automated decision makers. Arguably, such algorithmic decisions may be substantively flawed and lack in transparency and accountability. They also confront us with the moral, social and psychological implications of allowing machines to exercise decisional authority over human beings. In response to these concerns, several instrument aim to ban or regulate automated decisions in certain contexts. Article 22 of the GDPR provides that:

> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
>
> 2. Paragraph 1 shall not apply if the decision:
>
> (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests; or
>
> (c) is based on the data subject’s explicit consent.
>
> 3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
>
> 4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests are in place.[91]

[91] GDPR, art. 22.

Parallel provisions, albeit with somewhat different balancing formulas, can be found in the Malabo Convention (a ban on important automated decisions based on certain personal information) [92] and the Council of Europe's Convention 108+ (a requirement of considering one's views in the context of significant automated decisions).[93] The White House Blueprint also seems to recognise the general contours of a right to opt out from automated decisions: "You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter".[94]

The AI act does not ban, in and of itself, automated decision making, but introduces a right to explanation – a heightened transparency requirement – with regard to some decisions of AI systems, including the automated decisions covered by article 22 of the GDPR. Article 86 of the AI Act provides as follows:

> 1. Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system listed in Annex III, with the exception of systems listed under point 2 thereof, and which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.
>
> 2. Paragraph 1 shall not apply to the use of AI systems for which exceptions from, or restrictions to, the obligation under that paragraph follow from Union or national law in compliance with Union law.

[92] Malabo Convention, art. 14(5)("A person shall not be subject to a decision which produces legal effects concerning him/her or significantly affects him/her to a substantial degree, and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him/her").

[93] Modernised Convention for the Protection of Individuals with regard to the Processing of Personal Data, 18 May 2018, art. 9(1), ETS 223 (hereinafter: Council of Europe's Convention 108+)("Every individul shall have a right: (a) not to be subject to a decision significantly affecting him or her based solely on an automated processing of data without having his or her views taken into consideration").

[94] White House Blueprint, p. 46.

> 3. This Article shall apply only to the extent that the right referred to in paragraph 1 is not otherwise provided for under Union law.[95]

A broad right to explanation can also be found in the Italian AI Law,[96] which also underscores that responsibility for decisions taken by the public administration and judiciary must remain in the hands of the human officials.[97] In the same vein, the Chinese Personal Information Protection Law provides for a right to explanation of automated decisions entailing significant impact on the rights and interests of individuals, but also for a right to reject such decisions.[98]

Support for limiting the operation of automated decision-making systems can also be derived from the EU Declaration, which protects the right of workers to obtain human oversight regarding important decisions,[99] and the CoE Framework Convention, which invokes values of dignity, autonomy and oversight.[100]

[95] AI Act, art. 86. Note

[96] Italian AI Law, art. 15(3)("Citizens must be able to request explanations about decisions significantly affecting them that involve artificial intelligence").

[97] Italian AI Law, art. 11(3)("Human oversight of decisions remains essential, and the final responsibility always lies with the public official",13(1)("Artificial intelligence systems may be used to support judicial activities, provided that the judge always retains responsibility for the decision").

[98] Personal Infromation Protection Law (China), 20 August 2021, art. 24, https://personalinformationprotectionlaw.com/ (hereinafter: Chinese PIPL)(" Where personal information processors use personal information to make automatic decision, the transparency of decision-making and the fairness and justice of the results shall be ensured, and shall not impose unreasonable differential treatment on individuals in terms of transaction price and other transaction conditions… Where automatic decision-making has a significant impact on individual's rights and interests, he/she has the right to require the personal information processor to give an explanation, and to reject the decision made by the personal information processor only through automatic decision-making").

[99] EU Declaration, para 6(e)("[we commit to] ensuring in particular that human oversight is guaranteed in important decisions affecting workers, and that workers are generally informed that they are interacting with artificial intelligence systems").

[100] CoE Framework AI Convention, art. 7-8.

III. The case for an international AI bill of human rights

Limits of existing standards

A careful review of the standard-setting instruments adopted in recent years suggests that, while they do afford new legal protections and policy commitments in respect of the rights of individuals adversely affected by the use of AI systems, they generally fall short of the type and level of protections afforded by international human rights law. Furthermore, existing norms of international human rights law do not adequately respond to many of the human rights challenges posed by AI systems. As a result, there appears to be room for new normative developments that would address current protection gaps and normative shortcomings.

International human rights law norms are characterized by four principal features: a) universality – they confer legal rights on every human being merely by virtue of their humanity; b) inalienability – human rights held by individuals can never be removed from them (although many human rights can be restricted on the basis of a three-part test – (1) restriction by law; (2) when doing so is justified; and (3) where it is necessary and proportionate);[101] c) elevated normative status – human rights enjoy a relatively high normative status, sometimes even one that is peremptory in nature (*jus cogen*s). This means that human rights laws often have constitutional-like status in both domestic and regional law. Moreover, in all contexts, a strong justification is required for limiting them or derogating from them;[102] d) states as duty holders – in international law, human rights establish primarily a legal relationship between states as duty holders and individuals and groups of individuals as right holders.[103] In recent years, there have been attempts to expand this legal framework and to develop for business entities a

[101] See e.g., John O'Manique, Universal and Inalienable Rights: A Search for Foundations, 12 *Human Rights Quarterly* (1990) 465, 473.

[102] See e.g., Andrea Bianchi, Human Rights and the Magic of Jus Cogens, 19 *EJIL* (2008) 491, 495.

[103] See e.g., Christian Tomuschat, *Human Rights: Between Idealism and Realsim* (3rd ed., 2014) 119.

responsibility to respect human rights as a matter of social expectations first,[104] and as a matter of law later.[105]

Another, potentially fifth feature, which can be associated with in existing international human rights law, is an intermediate level of specificity. Key human rights instruments, like the Universal Declaration, ICCPR, ICESCR and the major regional human rights treaties, typically list between ten to thirty substantive provisions, each articulating a distinct human right. Such provisions often delineate with some degree of precision the scope of application of the covered right, enumerate its main elements, describe the correlative obligations of the duty holders and provide the conditions under which the enjoyment of the right may be restricted.[106] Only rarely they go beyond that and specify technical matters relating to their implementation.[107]

When viewed against these typical features, it is clear that most, if not all of the standard-setting instruments adopted in recent years which address, directly or indirectly, the human rights implications of AI systems are not formulated like typical

[104] UNGPs, p. 14.

[105] https://www.business-humanrights.org/en/big-issues/governing-business-human-rights/un-binding-treaty/.

[106] See e.g., ICCPR, art. 19 ("1. Everyone shall have the right to hold opinions without interference. 2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice. 3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals"); American Convention on Human Rights, 22 Nov. 1969, art. 5, 1144 UNTS 143 (hereinafter: American Convention)("1. Every person has the right to have his physical, mental, and moral integrity respected. 2. No one shall be subjected to torture or to cruel, inhuman, or degrading punishment or treatment. All persons deprived of their liberty shall be treated with respect for the inherent dignity of the human person. 3. Punishment shall not be extended to any person other than the criminal. 4. Accused persons shall, save in exceptional circumstances, be segregated from convicted persons, and shall be subject to separate treatment appropriate to their status as unconvicted persons. 5. Minors while subject to criminal proceedings shall be separated from adults and brought before specialized tribunals, as speedily as possible, so that they may be treated in accordance with their status as minors. 6. Punishments consisting of deprivation of liberty shall have as an essential aim the reform and social readaptation of the prisoners"); African Charter of Human and Peoples' Rights, 27 June 1969, art. 16, 21 ILM (1982) 58 (hereinafter: African Charter) ("Every individual shall have the right to enjoy the best attainable state of physical and mental health. 2. States parties to the present Charter shall take the necessary measures to protect the health of their people and to ensure that they receive medical attention when they are sick").

[107] See e.g., International Convention on the Protection of All Persons from Enfroced Disappearance, 20 Dec. 2006, art. 17(3), UN Doc. A/RES/61/177 (2007).

human rights instruments. Although they are useful in constraining power and protecting individual needs and interests, their ability to provide individuals potentially impacted by the use of AI systems effective human rights protections, comparable in kind and quality to that afforded by human rights treaties like the ICCPR or ICESCR, is rather limited. Arguably, the gap between current normative standards governing the use of AI systems and the protective framework normally associated with international human rights law could be filled by a new international AI bill of rights.

Universality

Some of the standard-setting instruments discussed above do not clearly afford universal protection to all human beings affected by AI systems. The most obvious example is the GDPR, which applies only in circumstances having an EU connection,[108] and does not use the language of universal human rights. Instead, it deals with the "rights of data subjects".[109] By contrast, certain standard-setting instruments, such as the Global Digital Compact,[110] EU Declaration,[111] the CoE Framework AI Convention[112] and the South Korean Basic AI Act,[113] refer explicitly to international or universal human rights. Other instruments are located somewhere in the middle of the spectrum between

[108] GDPR, art. 3 ("1. This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not. 2. This Regulation applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union, where the processing activities are related to: (a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or (b) the monitoring of their behaviour as far as their behaviour takes place within the Union. 3 This Regulation applies to the processing of personal data by a controller not established in the Union, but in a place where Member State law applies by virtue of public international law"). See also AI Act, art. 2(1)("This Regulation applies to: (a) providers placing on the market or putting into service AI systems or placing on the market general-purpose AI models in the Union, irrespective of whether those providers are established or located within the Union or in a third country; (b) deployers of AI systems that have their place of establishment or are located within the Union; (c) providers and deployers of AI systems that have their place of establishment or are located in a third country, where the output produced by the AI system is used in the Union; (d) importers and distributors of AI systems; (e) product manufacturers placing on the market or putting into service an AI system together with their product and under their own name or trademark; (f) authorised representatives of providers, which are not established in the Union; (g) affected persons that are located in the Union").

[109] GDPR, Ch III (Rights of the data subject).

[110] Global Digital Compact, para. 22-25.

[111] EU Declaration, premable.

[112] CoE Framework AI Convention, art. 4.

[113] South Korean Basic AI Act, art. 1.

particularity and universality: The AI Act uses the language of "fundamental rights"[114] – the EU law parallel to human rights – and the White House Blueprint and Malabo Convention refer to human rights only indirectly by alluding to formulations like "civil rights"[115] and the rights of "persons".[116] Finally, it should be noted that some regional human rights instruments underscore their interest in promoting values that have particular resonance in the regional context[117] – suggesting that they do not aspire to fully reflect universal standards.

Inalienability

The inalienability of rights is closely linked to the legal status of the norms introduced by the relevant standard-setting instruments. Declarations that do not have a binding legal effect – such as the EU Declaration or the Global Digital Compact – cannot confer, in and of themselves, inalienable rights. It is notable, however, that even some of the instruments that create legally enforceable rights or call for reinterpreting existing laws in accordance with the legal norms they introduce, appear to provide generous waiver provisions – a feature that appears to stand in tension with any presumptive status as them being inalienable in nature. For example, the right conferred by article 22 of the GDPR (not to be subject to automated decisions) hinges on lack of consent by the data subject[118] In the same vein, the White House Blueprint speaks of an opt out right from an automated service to a human alternative,[119] and the AI Act allows for informed consent to participate in the testing of high risk AI systems.[120] While having a human right typically includes an implicit right of choice as to whether or not to exercise the right, the explicit reference to the elective character of AI-related rights in the said standard-setting instruments may suggest that intended to introduce rights more tentative in nature than other human rights, such as freedom of expression and the right

[114] AI Act, recital para. 1. See also Italian AI Law, art. 4(1).

[115] White House Blueprint, p. 2.

[116] Malabo Convention, art. 14(5).

[117] AU Continental AI Strategy, p. 4.

[118] GDPR, art. 22(2) ("Paragraph 1 shall not apply if the decision:… (c) is based on the data subject's explicit consent"). But see, Malabo Convention, art. 14(5)(lacking an explicit consent exception).

[119] White House Blueprint, p. 46.

[120] AI Act, art. 60(4)(i).

to liberty, whose respective treaty provisions do not mention the possibility of waiver. This nuance is especially significant in light of the increasingly accepted view that, in digital contexts, consent requirements often exist only *pro forma*, and that they do not afford an adequate level of protection for digital rights.[121]

Elevated status

The same feature mentioned above – that many of the standard-setting instruments are non-binding in nature (and constitute soft law) – also renders it difficult to assign to them the elevated normative status that is characteristically associated with human rights norms. Yet, even binding instruments, such as the GDPR, AI Act or Malabo Convention, do not appear to have been intended to create norms that enjoy a high legal status comparable to human rights norms. The AI Act is not formulated at all in the language of rights, but rather as a set of largely technical requirements, and both the GDPR and Malabo Convention that do use a language of rights, resort to formulations that differ from those used in the regional human rights instruments of the regions in which they apply.[122] Significantly, none of those instruments were integrated within the institutional framework of regional human rights courts and commissions through protocols to existing human rights treaties.

It is also notable that both the GDPR and the AI Act provide generous exceptions to the rights they introduce – including legislative exceptions,[123] necessity exceptions[124] and security exceptions[125] – the latter suggesting that their scope of application does not

[121] See e.g., Neil Richards and Woodrow Hartzog, The Pathologies of Digital Consent, 96 *Washington University Law Review* (2019) 1461.

[122] *Cf.* Charter of Fundamental Rigths of the Euroepan Union, 18 Dec. 2000, art. 8, OJ C 364/1 (hereinafter: EU Charter)("Everyone has the right to the protection of personal data concerning him or her"); African Charter, art. 9(2)("Every individual shall have the right to express an disseminate his opinions within the law").

[123] GDPR, art. 22(2)(a)("Paragraph 1 shall not apply if the decision:… (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests").

[124] GDPR, art. 22(2)("Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller".

[125] GDPR, art. 2(2)("2. This Regulation does not apply to the processing of personal data:… (b) by the Member States when carrying out activities which fall within the scope of Chapter 2 of Title V of the TEU"); AI Act, art. 2(3)("3. This Regulation does not apply to areas outside the scope of Union law, and shall not, in any event, affect the competences of the Member States concerning national security, regardless of the type of entity entrusted by the

encompass some of the most important realms of governmental activity which could result in human rights infringements. Such wide protection gaps – which do not normally appertain to recognized human rights – do sit well with the proposition that these standard-setting instruments enjoy a normatively elevated status.

Some standard-setting instruments, however, do appear to have been intended to generate norms of elevated status. These include norms found in digital charters that purport to interpret or influence the interpretation of existing domestic constitutional law (e.g., the White House Blueprint and the Spanish Digital Rights Charter) and the two Council of Europe Conventions, which use the language of human rights,[126] and aim to strengthen international human rights protections.[127] The Council of Europe Conventions, however, have not been incorporated in existing human rights mechanisms and would – once in force – be monitored by separate new committees.

State obligations

Most of the standard-setting instruments alluded to in this White Paper are addressed to states and aim to empower individuals who may be impacted by AI systems. In this respect, they follow the general contours of international human rights law. Still, there are some exceptions to this state-centred configuration. The EU regulations discussed above involve states and individuals, but actually focus on the relationship between private actors – AI developers or deployers, on the one side, and individual users or

Member States with carrying out tasks in relation to those competences. This Regulation does not apply to AI systems where and in so far they are placed on the market, put into service, or used with or without modification exclusively for military, defence or national security purposes, regardless of the type of entity carrying out those activities. This Regulation does not apply to AI systems which are not placed on the market or put into service in the Union, where the output is used in the Union exclusively for military, defence or national security purposes, regardless of the type of entity carrying out those activities").

[126] Convention 108+, art. 9(1)("Every individual shall have a right…"; note, however, that the title of article 9 is "Rights of the data subject"); CoE Framework AI Convention, art. 4 ("Each Party shall adopt or maintain measures to ensure that the activities within the lifecycle of artificial intelligence systems are consistent with obligations to protect human rights, as enshrined in applicable international law and in its domestic law").

[127] Convention 108+, art. 1 ("The purpose of this Convention is to protect every individual, whatever his or her nationality or residence, with regard to the processing of their personal data, thereby contributing to respect for his or her human rights and fundamental freedoms, and in particular the right to privacy"); CoE Framework AI Covnention, art. 1(1)("The provisions of this Convention aim to ensure that activities within the lifecycle of artificial intelligence systems are fully consistent with human rights, democracy and the rule of law.").

data subjects, on the other hand. For instance, the AI Act prohibits “the placing on the market, the putting into service or the use” of certain AI systems,[128] and requires high-risk AI systems to comply with certain requirements. In the same vein, the EU Digital Services Act requires that “[p]roviders of very large online platforms and of very large online search engines shall diligently identify, analyse and assess any systemic risks”.[129] While international human rights law also imposes a positive duty on states to protect the rights of one person against infringements by other natural or legal persons, these AI regulating instruments resemble more in their structure and contents consumer protection laws that create rights for consumers vis-à-vis commercial actors they engage with, than human rights norms that establish a direct right of claim for individual victims against the state (and, perhaps, others sources of power and authority). The White House Blueprint is another example of an instrument that appears to be more focused on regulating the conduct of “[d]esigners, developers, and deployers of automated systems”[130] that on regulation of the conduct of federal government itself – including regulating its regulatory powers.

To be sure, addressing relationships between private actors is critically important for the effective implementation of digital human rights given the central role played by AI and other technology companies in this space. Still, the combined effect of instruments not using the language of human rights and not focusing on the traditional relations between states and individuals, complicates our ability to treat them as comparable to other international human rights instruments.

Level of specificity

Many of the reviewed standard-setting instruments differ significantly in their level of specificity from international human rights treaties, suggesting that there still is room for additional intermediate-level regulatory intervention. Not surprisingly, declarations like the Global Digital Compact or the AU Continental AI Strategy tend to allude to human

[128] AI Act, art. 5(1).
[129] DSA, art. 34(1).
[130] See e.g., White House Blueprint, p. 40.

rights in abstract and general terms (see e.g., “[t]o achieve our goal, we will pursue the following objectives:… Foster an inclusive, open, safe and secure digital space that respects, protects and promote human rights”[131] and “[t]he production, development, use and assessment of AI systems in Africa will always uphold human dignity, gender equality and respect and promote all the human rights set out under the African Charter on Human and Peoples’ Rights and its subsidiary instruments, as well as the Universal Declaration on Human Rights and related instruments of international human rights law”[132]). Yet, even the Council of Europe Framework AI Convention – which will eventually become a biding legal instrument – uses open-ended formulations, which provide state parties relatively limited guidance:

> Article 7 – Each Party shall adopt or maintain measures to respect human dignity diversity, fairness, social justice, and internationally recognized labour rights”, and individual autonomy in relation to activities within the lifecycle of artificial as well as to principles such as transparency, accountability and safety; the intelligence systems.
> Article 8 –Each Party shall adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place in respect of activities within the lifecycle of artificial intelligence systems, including with regard to the identification of content generated by artificial intelligence systems.
> Article 9 –Each Party shall adopt or maintain measures to ensure accountability and responsibility for adverse impacts on human rights, democracy and the rule of law resulting from activities within the lifecycle of artificial intelligence systems.
> Article 10 – Each Party shall adopt or maintain measures with a view to ensuring that activities within the lifecycle of artificial intelligence systems respect equality, including gender equality, and the prohibition of discrimination, as provided under applicable international and domestic law. Each Party undertakes to adopt or

[131] Global Digital Compact, para. 7. See also the 2019 OECD Council Recommendations on AI https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (the Recommendations were amended in 2024).

[132] AU Continental AI Strategy, p. 28.

> maintain measures aimed at overcoming inequalities to achieve fair, just and equitable outcomes, in line with its applicable domestic and international human rights obligations, in relation to activities within the lifecycle of artificial intelligence systems.
> Article 11 –Each Party shall adopt or maintain measures to ensure that, with regard to activities within the lifecycle of artificial intelligence systems: a b privacy rights of individuals and their personal data are protected, including through applicable domestic and international laws, standards and frameworks; and effective guarantees and safeguards have been put in place for individuals, in accordance with applicable domestic and international legal obligations.

It is notable that the drafters of the Framework Convention did consider complementing its general text with a detailed list of rights.[133] Some participants in the consultations conducted as part of this research programme, who have been involved in different stages of these negotiations, suggested that agreement around such a detailed list was not deemed political feasibility at the time and that the initiative of compiling a detailed list of relevant human rights was consequently dropped.

On the other side of the specificity-generality spectrum, one can find very detailed instruments, such as the AI Act, that provide elaborated guidance on the use of specific AI systems, without always clearly articulating which, if any, human right they implicate. For example, article 5(1)(a) of the AI Act prohibits “the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person’s consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm”. Another example from the same article involves a lengthy

[133] Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI), Feasibility Study, 17 Dec. 2020, CoE Doc. CAHAI(2020)23, p. 27 *et seq*

and complex technical arrangement concerning the use of ‘real-time’ remote biometric identification systems[134] Yet another detailed arrangement exists for post-remote biometric identification systems.[135]

At the same time, a few standard-setting instruments are formulated in ways that mirror in their level of specificity human rights treaties. These include article 22 of the GPDR, Convention 108+, the Malabo Convention, the EU Declaration and the White House Blueprint. Still, as explained above, these instruments tend to differ in other critical ways from human rights treaties.[136]

Limits of existing human rights norms

One key question regarding the need for a new international AI bill of human rights is whether existing human rights could effectively address the basic needs and interest of individuals impacted by AI systems. Obviously, there is no need for a new bill of human rights, if existing human rights cover new needs and interests adequately. Such a sentiment is often linked to fears of “right inflation” – that is, concerns that a sharp increase in the number of rights will dilute the value of all rights.[137] It is also linked to fears of acknowledging normative vacuums – that is, that the very attempt to establish new human rights might expose the existence of normative gaps and hamper our ability to invoke existing human rights in the interim, before new human rights are legislated.

A review of the human rights challenges posed by the AI systems which were considered in Part One and which the new standards setting instruments discussed in Part Two tried to address, suggests that existing international human rights law norms can protect many, but not all of the basic needs and interests implicated by AI systems.

[134] AI Act, art. 5(1)(h), 5(2)-(7).

[135] AI Act, art. 26(10).

[136] As explained before, article 22 of the GDPR protects the rights of data subjects, and the EU Delcaration is not intended to create binding norms. The White House Blueprint is not binding, focuses on US civil rights (and not interntaional human rights) and is formulated in language that is deliberately colloquial in style. Convention 108+ (which is not force yet) and Malabo Convention are formulated in ways that mirror human rights treaties, but address AI human rights to a limited extent only.

[137] See e.g., Jens T. Theilen, The Inflation of Human Rights: A Deconstruction, 34 *Leiden Journal of International Law* (2021) 831.

Furthermore, some existing human rights provide only partial coverage to such needs and interests, and lack the degree of specificity required to afford them with effective levels of protection. As a result, it might be desirable to build upon existing human rights and complement them with new and more detailed human rights norms – as has been done throughout the history of international human rights law when new needs and interests that were not fully, clearly and effectively protected by pre-existing human rights were identified. This approach was accepted, in principle, by most participants in the consultations (although some were more sceptical than others about the feasibility of legislating new human rights instruments at this point in time).

Which existing rights can be applied to AI systems?

Two human rights which are clearly relevant for the use of AI systems are the right to privacy and the right to non-discrimination. The former right can address the heightened privacy risks posed by AI systems due to their potential for privacy infringement through processes such as decryption, de-anonymization, biometric recognition and data-mining, and because of the insatiable appetite of AI systems for new data. The latter right can address specific problems stemming from biases in training and post-training data, discriminatory data labelling and algorithmic design, as well as problems associated with the use of proxies, the 'tyranny of metrics'[138] and the impact of all these features on the perpetuation of existing societal inequalities.

But even when pre-existing rights are applicable, many open questions remain regarding the scope of application of these rights, their precise manner of implementation in the context of the use of AI systems and the balance to be struck against competing rights and interests. The lingering effect of these questions could justify the elaboration of more specific human rights standards to address them. Such standards may include, with respect to the right to privacy for example, questions relating to the operationalization of information self-determination (i.e., the right of individuals to exercise control over data in the hands of data controllers and data

[138] See e.g., Jerry Z. Muller, *The Tyranny of Metrics* (2019).

processors),[139] privacy by design obligations (including the introduction, when appropriate, of techniques such as differential privacy)[140] and conditions under which consent can be meaningfully granted in digital contexts. With respect to non-discrimination, more detailed human rights norms could provide specific guidance on issues such as algorithmic fairness,[141] disparity testing[142] and the use of reasonable inferences.[143]

In addition, some others aspects of AI human rights can be read into existing human rights instruments with relative ease, by regarding such AI rights as relevant conditions for enjoying pre-existing rights. This is particularly the case with regard to access to AI systems, as well as quality requirements associated with such access – e.g., safety, reliability and effectiveness. One can argue, for example, that the right to life and the right to health demands quality access to AI systems that can support life-saving and health-enhancing treatment, and that in the same vein, the right to education demands quality access to educational AI. It can also be argued that quality access to AI systems is required by virtue of the ICESCR's right to enjoy the benefits of scientific progress.[144] Still, here too, further clarity regarding the scope of these interpretations, the constitutive elements of the AI human rights that have been read into pre-existing rights (e.g., safety, reliability, effectiveness) and the limitations attached thereto (e.g., financial limitations, unacceptable risks) could be highly beneficial.

One example of the problem of partial coverage can be found with respect to algorithmic transparency. While traditional human rights law does not include

[139] See e.g., Florent Thouvenin, Informational Self-Determination: A Convincing Rationale for Data Protection Law?, 12 *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* (2021) 246.

[140] See e.g., Cynthia Dwork and Aaron Roth, The Algorithmic Foundations of Differential Privacy, 9 *Theoretical Computer Science* (2014) 211.

[141] See e.g., Xiaoment Wang, Yishi Zhang and Ruilin Zhu, A Brief Review on Algorithmic Fairness, 1 *Management System Engineering* (2022) 7.

[142] See e.g., Harvineet Singh and Rumi Chunara, Measures of Disparity and their Efficient Estimation, *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023) 927.

[143] See e.g., Sandra Wachter and Brent Mittelstad, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, 2019 *Columbia Business Law Review* 494.

[144] ICESCR, art. 15(1)(b)("[The States Parties to the present Covenant recognize the right of everyone:] To enjoy the benefits of scientific progress and its applications").

transparency as a standalone right, it can be regarded as a procedural requirement needed to ensure lack of arbitrariness in official decisions, especially those involving the right to life, liberty and privacy.[145] It can also be regarded, in certain contexts, as part of the right to a "fair and public hearing"[146] and the right to seek and receive information.[147] Still, outside these particular legal contexts, there does not appear to be a general basis in existing international human rights law for a broad right to transparency, nor is there a clear basis for elaborating the different potential components of the right in the AI context – explainability/interpretability,[148] traceability[149] etc. – and for balancing the right against competing rights and interests (e.g., privacy and intellectual property).

In a similar vein, concerns about AI manipulation can be related, in certain cases, to existing rights to freedom of thought[150] and opinion,[151] and to other rights whose enjoyment can be undermined by way of manipulation, such as the right to health,[152] privacy,[153] personal security,[154] participation in public life,[155] and the right to property.[156] Still, the ability to identify, in certain cases, a relevant pre-existing human rights norm would not resolve the need to clearly define what constitutes manipulation (addressing, for instance, impersonation, use of synthetic contents, and exploitation of vulnerabilities and cognitive biases), identifying problematic manipulation techniques (e.g., subliminal messaging, emotion detection) and distinguishing them from other legitimate forms of influence.

[145] ICCPR, art. 6, 9, 17.
[146] ICCPR, art. 14(1).
[147] ICCPR, art. 19(2).
[148] See e.g., David A. Broniatowski, Psychological Foundations of Explainability and Interpretability in Artificial Intelligence, NISTIR 8367 (2021).
[149] See e.g., Marçal Mora-Cantallops, Salvador Sánchez-Alonso, Elena García-Barriocanal and Miguel-Angel Sicilia, Traceability for Trustworthy AI: A Review of Models and Tools, 5(2) *Big Data and Cognitive Computing* (2021) 20.
[150] ICCPR, art. 18(1).
[151] ICCPR, art 19(1).
[152] ICESCR, art. 12.
[153] ICCPR, art. 17.
[154] ICCPR, art. 9(1).
[155] ICCPR, art. 25.
[156] Protocol to the Convention for the Protection of Human Rights and Fudndamental Freedoms, 20 March 1952, art. 1, ETS 9; American Convention, art. 21; African Charter, art 14; Universal Declaration of Human Rights, 10 Dec. 1948, art. 17, UN Doc. A/RES/217 (iii).

Finally, concerns about lack of accountability for the use of AI systems do coincide with the pre-existing right to effective remedy[157] and with the states' duty to ensure human rights protection.[158] In the absence of an accountable legal or natural person, the ability of victims of violations of existing human rights that are facilitated by the use of AI systems to obtain *ex ante* effective protection and *ex post* remedy is seriously compromised. The modalities for holding individuals and entities involved in the development, dissemination and use of AI systems would need, however, to be clearly spelled out in ways that would minimize accountability gaps related to the private and transnational features of AI technology, and in a manner that would strike a reasonable balance between accountability and the pragmatic need for "safe harbours",[159] and fair and predictable application of liability laws.

Which new rights should be developed?

There are some aspects of the use of AI systems which impact basic human needs and interests that could be regarded as giving rise to strong claims for new human rights protections due to their critical effect upon personal wellbeing, their close relationship to moral principles underlying many human rights such as dignity, liberty and solidarity, and the lack of adequate coverage by pre-existing human rights. These needs and interests are threatened by failures to ensure adequately regulated access to AI systems, AI bias and fairness, AI transparency, protection from AI manipulation and AI accountability.

In particular, there may be a need for protecting individuals from being made subject to automated decisions and to algorithmic interactions *in lieu* of human-to-human interaction.[160] These latter needs and interests are only partly protected by existing human rights law (e.g., in contexts related to legal proceedings, where human

[157] ICCPR, art. 2(3).

[158] ICCPR, art. 2(1).

[159] See e.g., Iskandar Haykel, The Stick, the Carrot, and the Net: Policy Approaches for Addressing AI Agent Harms (2025), https://ari.us/wp-content/uploads/2025/08/AI-Liability-Report-The-Stick-the-Carrot-and-the-Net.pdf.

[160] For a discussion, see Yuval Shany, A Right to Human-to-Human Interaction, 7 Nov, 2024, *AI Ethics at Oxford Blog*, https://www.oxford-aiethics.ox.ac.uk/blog/right-human-human-interaction.

involvement may be regarded part of applicable due process guarantees[161] and, in humiliating circumstances, as part of the right not to be subject to inhuman or degrading treatment).[162] As a result, it appears that only the introduction of a new human right to a human decision and interaction would give effective protection to the full gamut of basic needs and interests impacted by automated decisions and interactions – e.g., requiring quality performance, fair and transparent process and dignified treatment – in all circumstances involving them.

As indicated above, during the consultations I held, there was broad support for the usefulness of elaborating human rights standards applicable to AI systems, given the lack of a human rights language specific to AI systems in existing human rights instruments, and the paucity of human rights-like intermediate level of regulation in standard-setting instruments dealing with to AI systems. Yet, there was a divergence of opinions among participants in the consultations regarding the existence of an imperative need for a new bill of human rights. Some were of the view that it is preferable to gradually build on existing legal standards through judicial interpretation and academic work (including through the development of detailed case studies addressing different aspects of AI and human rights) and through the development of sectoral approaches. Such a divergence of positions can be explained, in part, by differences of opinion among participants regarding the desirability of investing judges and other legal experts with developing new human rights law (something that raises democratic legitimacy concerns),[163] and the likelihood of developing new standards in the field in the current geo-political context which appears to be dysfunctional, national security-oriented and anti-regulatory in its orientation.

To my mind, the better view is that human rights protections are crucially and urgently needed in order to steer and critically evaluate the fast development of AI systems and

[161] ICCPR, art. 9, 14.

[162] ICCPR, art. 7.

[163] For a discussion, see Yuval Shany, The Democratic Legitimacy of Digital Human Rights, in EE Research Handbook on International Economic Law and Human Rights (Holger Hestermeyer and Carla Lopez eds., forthcoming in 2026).

their impact of human wellbeing and human rights. Such effective human rights protections are currently unbailable due to insufficiently precise nature and partial coverage of pre-existing human rights standards. The need for undertaking a conscious effort to elaborate new human rights norms – which would largely build on pre-existing human rights norms – is particularly important given the paucity of judicial interventions in this policy space. This is partly due to the limited role of states, the traditional duty holder in international human rights law, in the development and dissemination of the technology. By introducing and elaborating new human rights norms, we would be affording individuals and groups of individuals potentially impacted by AI systems a powerful vocabulary to articulate their needs, interests and demands in the language of human rights. This could facilitate efforts to push tech companies to align their AI systems with international human rights standards, specifically tailored for such systems, in ways that would complement parallel initiatives to promote alignment with ethical standards (which are often too vague and subjective).

Furthermore, the more democratic the process of new norm-creation and adaptation will be, the more legitimate are calls on all actors – including private actors – to follow the ensuing norms. From this vantage point, there is a clear advantage for developing new human rights norms through democratic-representative bodies rather than through judicial or expert interpretations. Such interpretations sometimes tends to over-extend existing human rights law norms – causing political back-lash – and to generate mega-rights (like the right to privacy) which are hard to delimit conceptually and to apply in practice.[164]

Ultimately, the same reasons that justify the development of, and reliance on human rights law in other areas of life – i.e., to protect and promote moral entitlements in situations where they are, or expected to be, under considerable pressure – apply to adapting human rights for use cases involving AI systems. International human rights

[164] For a discussion, see Yuval Shany, How far can you go? Shoehorning Digital Rights into Existing Human Rights Treaties, *Vienna Journal on International Constitutional Law* (forthcoming in 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5180289.

law offers a set of recognized principles with universal applicability and powerful normative resonance (reflected, *inter alia,* in their inalienability and elevated legal status) which constrain power and protect individual needs and interests. International human rights law has already internalized the important insight that power and authority that merits constraining and channelling in ways compatible with the wellbeing of individuals and groups is not placed exclusively in the hands of states. As a result, business entities, including AI companies and other private actors using AI produces and services, should be subject to state regulation and international regulation under the UN Guiding Principles 'duty to protect' pillar.[165] AI companies should also be expected themselves to respect human rights[166] and both states and companies should facilitate access to remedies to victims of human rights violations and abuses.[167] One of the potential advantages of an international AI bill of human rights is that it could further clarify and consolidate the relevant normative expectations not only from states but also from AI companies and international organizations – whose role in this sphere may be significant given the need to coordinate between different national and regional regulatory schemes, and to develop a united front against big tech companies.

[165] UNGPs, Pillar I.
[166] UNGPs, Pillar II.
[167] UNGPS, Pillar III.

IV. The way forward: Towards a new international AI bill of human rights?

A soft law instrument

A key question that arose in the consultations is how to proceed in the direction of a new international AI bill of human rights, if one were inclined to do so (an option that many, though not all, participants expressed an interest in exploring). There was a broad consensus among participants that a global treaty is unlikely to be concluded on this topic, and even were it to be adopted, it would likely have highly diluted contents (like the Council of Europe Framework AI Convention). Such pessimism may reflect both global tensions and the hostility of the current USA administration toward most forms of AI regulation.[168] Only slightly little more optimism was expressed during the consultations with regard to the prospects of concluding an AI protocol to one or more of the regional human rights treaties.

The course of action deemed to be most realistic by many consultation participants was to aim first for the adoption of a soft law instrument listing the different AI human rights and enumerating their contents and manner of application, without addressing in detail the institutional framework that would be invited to apply them in due course (something that would need to be agreed upon at a later stage). Such a document could influence the legal interpretation of existing human rights instruments and serve as the basis for future law-making once conditions for moving in the direction of adopting binding law would ripen. A soft law instruments could be adopted, relatively quickly however, by international organizations operating at the regional or global level, by professional associations, civil society groups and/or academic institutions.

The history of the Universal Declaration of Human Rights itself – including the 1940 H.G. Wells and Sankey Declaration of the Rights of Man,[169] the 1946 American Law

[168] See e.g., Jeffrey Dastin and Ingrid Melander, Vance tells Europeans that heavy regulation could kill AI, *Reuters* 11 Feb. 2025, https://www.reuters.com/technology/artificial-intelligence/europe-looks-embrace-ai-paris-summits-2nd-day-while-global-consensus-unclear-2025-02-11/.

[169] A Declaration of the Rights of Man - A charter prepared in 1940, under the Chairmanship of Lord Sankey, and originally drafted for discussion by H. G. Wells, http://www.voting.ukscientists.com/sankey.html.

Institute Draft (prepared by Alvaro Alvarez),[170] the 1947-1948 UNESCO Survey[171] and the 1946-48 process of negotiations in the UN Human Rights Commission and the Third Committee[172] offer a (high profile) case study in this regard: Work on it started during World War Two, at a time when the prospects for reaching international consensus on sensitive matters with strong ideological dimensions appeared to be very slim. The drafting process involved professional associations, informal groups, and a number of international agencies, and also included extensive consultations with states, experts and professional associations. It resulted in a soft law document, which later became the basis for hard law – the 1966 Covenants on human rights.

Another relevant example is the gradual development of norms in the area of business and human rights – a consensus building exercise directed by John Ruggie, the UN Secretary General's Special Representative. The process resulted in the Ruggie Principles (2008),[173] which later became the UN Guiding Principles on Business and Human Rights (2011). The Guiding Principles, in turn, have influenced domestic and regional legislation – including the recent EU Corporate Sustainability Due Diligence Directive – and underlie efforts (so far, futile in nature) to conclude a legally binding instrument on the same topic.[174] An international AI bill of human rights could follow a comparable path of gradual legal development.

Some of the standard-setting instruments discussed here above also offer possible modalities for an AI Bill of Rights. The White House Blueprint is a particularly interesting example which could perhaps be replicated at the international level. The Blueprint,

[170] American Law Institute, Statement of Essential Human Rights, by a Committee Appointed by the American Law Institute, 243 *The Annals of theAmerican Academy of Political and Social Science* (1946) 18.

[171] UNESCO (ed.), *Human Rights: Comments and Interpretations* (a symposium with an introduction by Jacques Maritain), UN Doc. UNESCO/PHS/3 (rev.)(1948).

[172] For a discussion, see William A. Schabas, Introductory Essay: The Drafting and Significance of the Universal Declaration of Human Rights, *The Universal Declaration of Human Rights: The Travaux Preparatoires* (William A. Schabas ed., 2013) lxxi.

[173] Report of the Special Representative of the Secretary-General on the issue of Human Rights and Transnational Corporations and other Business Enterprises, Protect, Respect and Remedy: a Framework for Business and Human Rights, UN Doc. A/HRC/8/5 (2008).

[174] https://www.business-humanrights.org/en/big-issues/governing-business-human-rights/un-binding-treaty/.

whose focus is on US public law, contains five rights – each articulated in an expansive manner (on average, 234 words per right), followed by a detailed policy rationale ("why this principle is important"), an explanation of the implications of the right for the development of technological standards and practices ("what should be expected of automated systems"), real life examples of implementation ("how these principles can move into practice") and an appendix with examples of automated systems. These latter parts of the Blueprint coincide with the expectation expressed by some participants in the consultation that the project will generate a set of case studies that exemplify the promise and shortcomings of pre-existing human rights approaches, the need for new human rights norms to fill existing protection gaps, and the imperative of operating alongside other regulatory instruments and legal norms.

Possible contents of an international AI bill of rights

This White Paper survey of human rights-related needs and interests (Part One), recent standard-setting initiatives (Part Two) and existing protection gaps (Part Three), draws attention to seven distinct areas in which further normative development may be warranted, and where such developments have already started to materialize. In some areas, hat is needed is merely the fine-tuning of pre-existing human rights norms and their adjustment to some specific challenges related to AI systems. Yet, in some other areas, the development of wholly new human rights ought to be considered.

Given the preference afforded in this White Paper (and in the consultations) to promoting a soft law instrument, the doctrinal question of introducing legally binding human rights obligations for non-state actors would be largely skirted at this stage. This should not be read, however, as a rejection of the normative proposition that private actors should, whenever possible, respect international human rights law and that legal norms should develop in the future to legally require them to do so.

In line with a comment made earlier in this White Paper, it is also important to ensure that the precise contours of human rights laws would be shaped through democratic deliberation processes. Such processes should delineate the scope of application of human rights norms, identify the legal obligations that should attach to them and strike a balance between AI human rights and competing rights and interests (i.e., undergo a process of democratic concretization).[175] The list of rights and aspects of rights provided hereby should not be seen as an attempt to pre-empt such a deliberative process. Rather, it should be regarded as an invitation to policy-makers and stakeholders to enter into a conversation about the contents of a future AI bill of rights.

Finally, the rights listed here should not be read as a closed list. Existing human rights laws continue to apply to uses and failures to use AI systems, as do other legal, administrative, ethical and technological safeguards. The list of proposed rights is merely designed to complement, not displace, all other applicable standards.

1. Right of Access to AI Systems:

Rationale – An international AI bill of human rights should clearly indicate that individuals have a right to access AI systems, when such systems are reasonably available and cleared for use generally or in specific sectors. Such a right reflects the growing importance of AI system for realizing human rights in a variety of fields of human activity (e.g., health, education, labour), and is indicative of growing concerns about discrimination and unfairness emanating from an 'AI-divide'[176] between the Global North and Global South, and – inside countries – between 'AI natives' and 'AI illiterates'.[177] It can be regarded as an AI-specific application of the (somewhat esoteric) right to enjoy the benefits of scientific progress,[178] as well as an element of all substantive human right implicated by lack of access (e.g., right to health, right to education, right to work). The more our societies rely on AI systems, the stronger is the

[175] For a discussion, see Samantha Besson, The Legitimate Authority of International Human Rights: On the Reciprocal Legitimation of Domestic and International Human Rights, in *The Legitimacy of International Human Rights Regimes: Legal, Political and Philosophical Perspectives* (Andreas Føllesdal, Johan Karlsson Schaffer and Geir Ulfstein eds., 2013) 32.

[176] See e.g., International Labor Otganization and UN Office of the Secretary General's Envoy on Technology, *Mind the AI Divide: Shaping a Global Perspective on the Future of Work* (2024)

[177] See e.g., Daisy Thomas, Becoming AI-Literate and AI-Native in a Rapidly Evolving World, *Medium* 10 Jan. 2024, https://medium.com/@daisygarciathomas/becoming-ai-literate-and-ai-native-in-a-rapidly-evolving-world-5e7066c566ba.

[178] ICSECR, art. 15(1)(b).

case for regarding access to such systems also as an element of adequate standard of living.[179]

One may note that the aforementioned right to enjoy the benefits of scientific progress has been understood to cover not only the right to consume products and services generated by scientific innovation, but also to actively partake in their creation, when possible, and in shaping the ways in which they are configured and applied in the world.[180] A similar claim can be made with regard to the right of individuals and groups of individuals to partake in the development and design of AI systems and in policy-making regarding their manner of use.

Scope – The right to access has both collective dimensions – especially at the technology development and dissemination stage – and individual dimensions, especially at the post-launch stage. In all cases, the right should include proactive aspects intended to facilitate ease of access – e.g., availability, affordability, openness, inter-operability, user-friendliness – as well as aspects relating to the quality of the AI systems to which access is facilitated. These quality conditions should reflect expectations that AI systems to which access is granted are safe, secure, reliable, trustworthy, resilient and effective, operate at good speed with good data infrastructures in place, and compatible with human rights in their operations (including the right to privacy and non-discrimination). Once several AI systems are reasonably available to offer any particularly service, the right to access should allow, when possible, individual choice between these different systems.

Limitations – The right to access is a relative right. The positive obligations associated with it are subject, like other positive obligations, to due diligence or reasonableness requirements, and, like other economic, social and cultural rights, they involve the principle of progressive realization (which, in turn, gives expression to expectations of

[179] ICESCR, art. 11.

[180] See e.g., Samantha Besson, The 'Human Right to Science' *qua* right to participate in science: The participatory good of science and its human rights dimensions, 28 *The International Journal of Human Rights* (2023) 497.

international assistance and cooperation).[181] Beyond budgetary considerations that may slow down the timeline for reaching universal access to AI systems, restrictions on access may also be warranted in certain cases on the basis of the three-part test that is found in many other human rights contexts (legality of the restrictive measures, its legitimate aim, and necessity and proportionality). In relation to aspects that are covered by the ICESCR, limitations must be made “only in so far as this may be compatible with the nature of these rights”.[182]

Limitations on access to AI systems may be justified, for example, by way of reference to the aforementioned conditions for their use (e.g., safety, reliability and effectiveness), in order to protect the rights of others (e.g., privacy or non-discrimination) and for protecting the public interest in areas where the use of AI systems may be particularly disruptive and feature unacceptable levels of societal risk (e.g., certain use of facial recognition or emotional detection technology).

2. Privacy-related protections from harmful uses of AI systems

Rationale – The use of AI systems creates new and enhanced privacy risks due to the combined effect of extensive data collection for both supervised and unsupervised machine learning processes, the capacity of AI systems to personalize information and to circumvent data protection safeguards (such as encryption and anonymization) and the ability to utilize these systems to access sensitive personal data, such as biometric data, emotional states and subliminal reactions. While many privacy protections are already afforded by the right to privacy, and in the EU context – also by the right to data protection[183] – a specific human right may be necessary to afford clear, comprehensive and effective privacy protection standards, specially tailored to the unique challenges posed by AI systems.

[181] ICESCR, art. 2(1).
[182] ICESCR, art. 4.
[183] EU Charter, art. 8.

Scope – The right to privacy-related protections against AI harms should encompass the full life cycle of AI systems, covering data collection, data retention, access to data, data analysis (including, inferring sensitive personal information), data transfers and system outputs. Special attention ought to be afforded to the need to ensure that AI systems are not used to dilute or circumvent altogether existing privacy regimes (including, those based on data minimization and special purpose limitations), to the need to protect sensitive personal information, to the need to respect informational self-determination and to concerns about the extraordinary and cumulative harms associated with continuous and ubiquitous surveillance. It is also imperative that privacy protections would be integrated, as far as possible, in the development of AI systems ('privacy by design'),[184] and that reliance on user consent should be made subordinate to considerations of avoiding serious harm. Reliance on consent should also be mindful of the shortcomings of online modalities for giving consent, including the ability of deployers of AI system to manipulate user into consenting.

Limitations – like other privacy protections, the right to privacy-related protections against harmful uses of AI systems may be restricted on the basis of the aforementioned three-part test, which invites consideration, inter alia, of the public benefits accruing from the development and use of AI systems for particular purposes. It should often be required, in line with the principle of proportionality, that alternative privacy safeguards would be introduced to compensate for any adverse effect of the applied restriction (e.g., access to training data should be conditioned, at times, on its anonymization and purpose limitation) and that the record of performance of privacy-restricting systems should be made subject to ongoing review and oversight. The use of AI systems for surveillance purposes should require a heightened level of justification, and some surveillance systems which are particularly comprehensive and intrusive, including systems involving the collection and processing of sensitive personal information that feed into systems of social control such as social scoring, should be banned altogether.

[184] See e.g., Woodrow Hartzog, *Privacy's Blueprint: The Battle to Control the Design of New Technologies* (2018).

3. The right to be free from algorithmic bias and unfairness

Rationale – The use of AI systems creates new risks for discrimination, including structural, indirect, statistical and intersectional discrimination, caused and made harder to identify and remedy by the systems' unique features. The combined effect of lack of transparency about data use and data processing, including the use of profiling and drawing inferences, the propensity of AI systems to perpetuate, and at time exacerbate pre-existing inequalities, and their capacity to circumvent safeguards against non-discrimination, render the use of AI systems particularly prone to discrimination. At a deeper level, the evaluation and treatment of individuals on the basis of quantitative criteria and group membership raises new problems of unfairness and exacerbates pre-existing problems of such a nature.

The prohibition of non-discrimination already exists in human rights law both as a standalone right and as a component of existing substantive rights, whose enjoyment must be afforded in a non-discriminatory manner.[185] Its inclusion in a future international AI bill of human rights is justified by the need to elaborate particular modalities of equality protections, including pro-active measures (e.g., technical de-biasing), that would constitute a specific application of the prohibition against discrimination to the particular challenge posed by the use of AI systems. What's more, the extensive reliance of AI systems on proxies requires an extension of non-discrimination laws, beyond 'suspect classes' with which discrimination is historically associated.[186] In the same vein, some implications of the use of AI systems that generate structurally unfair results on a society-wide level – such as perpetuation of existing socio-economic stratification – may go beyond the pre-existing scope of anti-discrimination norms in international human rights law.

Scope – The right not to be subject to algorithmic bias and unfairness should cover, like other AI human rights, the entire life cycle of AI systems. It should cover biases in the system generated through the collection and labelling of training data as well as post-

[185] See e.g., ICCPR, art. 26.

[186] See e.g., ICCPR, art. 2(1), ICESCR, art. 2(2).

training data used for machine learning, flaws in the algorithmic design and in its application to particular tasks to which it is unsuitable. The right should also cover possible remedial measures designed to increase the fairness of AI systems, including through the development of corrective algorithms (e.g., ‘debiasing by design’), the introduction of guardrails in respect of problematic tasks that may be assigned to AI systems and the conduct of performance evaluation and disparity testing in the development and post-deployment stages.

In terms of types of discrimination that should be addressed, the right should cover both direct and indirect forms of discrimination[187] (keeping in mind that indirect discrimination, based on seemingly innocuous classifications, is expected to be common during the use of AI systems), disparate impact[188] or statistical discrimination[189] (whose probative value increases under conditions of non-transparency), intersectional discrimination[190] (whose prevalence is greatly enhanced by the vast number of factors used in algorithmic processing) and structural discrimination[191] (which overlaps to some extent with problems of algorithmic unfairness, manifesting themselves through the creation and perpetuation of societal structures that have disparate effects and non-equitable distributive implications across different population groups).

Particular attention should also be paid to the erosion or circumvention altogether of pre-existing non-discrimination safeguards, such as bans against invoking ‘suspect classifications’ or diversification of decision-making bodies (whose decision-making authority might have been now delegated to AI systems). Other aspects of non-

[187] See e.g., Sandra Fredman, *Discrimination Law* (3rd ed., 2022) 247 *et seq*.

[188] See e.g., Richard A. Primus, Equal Protection and Disparate Impact: Round Three, 117 *Harvard Law Review* (2003) 494.

[189] See e.g., Andra Moro, Statistical Discrimination, in *The New Palgrave Dicitionary of Economics* (Steven N. Durlauf and Lawrence E. Blume eds., 2009), https://link.springer.com/rwe/10.1057/978-1-349-95121-5_2972-1#citeas.

[190] See e.g., Kimberle Crenshaw, Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, 8 University of Chicago Law Forum (1989) 139.

[191] See e.g., Tom R. Burns, Towards a Theory of Strucutral Discrimination: Cultural, institutional and Interactional Mechanisms of the “European Dilemma”, in *Idenitity, Belonging and Migration* (Gerard Delanty, Ruth Wodak and Paul Jones eds., 2011) 152.

discrimination involve access to AI systems – especially by digitally illiterate persons and persons with disabilities – and the ability of AI systems to interact with users in a variety of minority and indigenous languages.

Limitations – The notion of algorithmic bias and unfairness must be informed by the foundational difference in non-discrimination laws between prohibited acts of discrimination and legitimate distinctions. In international human rights law, this difference is typically evaluated on the basis of the objectivity of the criteria used for applying different treatment to different individuals, their reasonableness and the proportionality of the harm the very use of the criteria and their real world implications entail, when compared to the societal benefits derived from the distinction.[192] These tests also invite the consideration of reasonable forms of accommodation and the imposition of proportionate burdens as potential fixes for problems of inequality.

In view of these considerations, the use of inferences and profiles by AI systems cannot always be considered unlawful. Furthermore, structural issues which raise fairness concerns can often be addressed only through incremental measures and/or affirmative action (sometimes referred to in international human rights law as temporary special measures),[193] which may generate, in and of themselves, questions regarding bias and fairness. In the same vein, closing gaps, like the 'AI divide', may require many progressive steps over time.

4. The right to algorithmic transparency and explainability

Rationale – The infamous 'black box' problem constitutes a defining feature of AI systems, whose seriousness appears to be growing the more AI systems exceed the data collection and data processing capacities of human beings. Basic differences in human and machine modes of "thinking" – one, based largely on causation, and

[192] See e.g., Antje von Ungern-Sternberg, Discriminatory AI and the Law: Legal Standards for Algorithmic Profiling, in *The Cambridge Handbook of Responsible Algorithmic Intelligence: Interdisciplinary Perspectives* (Silja Voeneky, Philipp Kellmeyer, Oliver Mueller and Wolfram Burgard eds., 2022) 252.

[193] See e.g., General recommendation No. 25, on article 4, paragraph 1, of the Convention on the Elimination of All Forms of Discrimination against Women, on temporary special measures, UN Doc. HRI/GEN/1/Rev.7 at 282 (2004).

another, based largely on correlation;[194] one centred on explanations and another on predictions[195] – further contribute to rendering the process by which machine outputs are generated beyond the cognitive reach of humans, with machine outputs often being non-interpretable and unpredictable even to developers of AI systems. This renders the exercise of decisional authority by way of reliance of AI systems seem, from the perspective of individual addresses of algorithmic decisions, arbitrary and unfair in nature. It also renders the performance of AI systems largely inscrutable to outside observers, including to those who seek to uphold their human rights vis-à-vis developers and deployers of AI systems. At a deeper level, lack of algorithmic transparency contributes to an epistemic crisis – a growing inability by individuals to understand the world around them, which may adversely affect their mental wellbeing and sense of agency. This may, in turn, result also in reduced capacity to hold public deliberations and in a democratic crisis.

There are some existing human rights norms that can be utilized for demanding algorithmic transparency and explainability. These include the ban on arbitrariness that attaches to certain human rights,[196] the requirement of due process, which typically includes a right to a reasoned and public judicial decision,[197] the right to take part in public life, which presumes a degree of public deliberation about public matters,[198] the right to seek and receive information, which may encompass information about algorithmic systems and decisions,[199] and the right to an effective remedy, which requires some information on decision-making processes leading to human rights

[194] See e.g., Thomas Hellström, The Relevance of Causation in Robotics: A review, Categorization and Analysis, 12 *Paladyn Journal of Behavioral Robotics* (2021) 238.

[195] See e.g., John Ball, Brains are not prediction machines, *Medium,* 2 April 2022, https://medium.com/pat-inc/brains-are-not-prediction-machines-a6983b04bc52.

[196] See e.g., ICCPR, art. 6, 9, 17.

[197] See e.g., ICCPR, art. 9(2), 14(1). Human Rights Committee, General Comment No. 35: Article 9 - Liberty and security of person, UN Doc. CCPR/C/GC/35, para. 25; Human Rights Committee, General Comment No. 32: Article 14 - Right to equality before courts and tribunals and to a fair trial, UN Doc. CCPR/C/GC/32 (2007) para. 49.

[198] See e.g., ICCPR, art. 25(1); Human Rights Committee, General Comment 25: Article 25 - Participation in Public Affairs and the Right to Vote U.N. Doc. CCPR/C/21/Rev.1/Add.7 (1996), para. 8 ("Citizens also take part in the conduct of public affairs by exerting influence through public debate and dialogue with their representatives or through their capacity to organize themselves. This participation is supported by ensuring freedom of expression, assembly and association").

[199] See e.g., ICCPR, art. 19(2).

violations.[200] In addition, the digital right to informational self-determination supports claims to receive information on personal data used by AI system and the manner by which such data was used.[201]

Still, certain aspects of the proposed right to algorithmic transparency and explainability remain under-protected by pre-existing human rights norms. It is highly debatable whether existing bans on arbitrariness and the right to seek and receive information capture the full gamut of uses of AI systems by private actors. Whereas transparency and reason-giving have long been considered important elements of good governance that could legitimate the exercise of public authority, they have not traditionally been regarded as human rights *per se*, and have not been applied, generally speaking, vis-à-vis private actors*.* Furthermore, it Is less than clear what specific duties of transparency and explainability could be reasonably expected from different actors across the value chain involved in the development and deployment of AI systems, and how should the right be balanced against competing considerations – most significantly, efficiency considerations and intellectual property concerns.

Scope – A right to algorithmic transparency should entail, arguably, certain disclosure obligations – first and foremost, a notification that an algorithm is being used and some basic information about the algorithm – the task for which it is used, the types of data it relies on, and the kinds of outputs it is expected to generate. This implies, *inter alia*, a duty to notify individuals interacting with AI system about the artificial nature of the system (with a view to avoiding deceptive impersonation of humans by AI systems) and the involvement of AI in the production of digital contents (with a view to avoid deceptive 'deep fakes' and provide information on other forms of synthetic contents).

In certain high risk contexts – for instance, when decisions with significant implications for individuals or groups of individuals are taken, or when safety breaches occur –

[200] See e.g., ICCPR., art. 2(3).

[201] See e.g., *Members of the José Alvear Restrepo Lawyers' Collective v. Colombia*, I/A CHR judgment of 18 Oct. 2023, para. 586; Case C-203/22, *CK v. Magistrat der Stadt Wien* (Dun & Bradstreet Austria GmbH case), CJEU judgment of 27 Feb. 2025, ECLI:EU:C:2025:117, para. 58.

heightened levels of disclosures would be warranted to relevant individuals, groups of individuals and/or their representatives. These may include technical information about the algorithm, the training and post-training data, processes of training, protocols for human-machine interfaces, impact assessment and oversight, and detailed information about the system's overall performance. Individuals should also be able to receive information on which of their personal data was used by AI systems and how such data was used. In addition, developers and deployers of AI systems should notify relevant members of the public about incidents that occurred during the use of AI systems and about vulnerabilities that were exposed.

Explanations regarding algorithmic decisions and recommendations subject to the right to explainability should be communicated, whenever possible, in a manner that is intelligible for laypersons, but also authentic in that they reflect the actual 'reasons' used in the data analysis process. They should also strive to afford traceability.[202] When heightened transparency and explanation-giving obligations attach, a technical examination of system performance by authorized monitors should be allowed with a view to attaining, if possible, interpretability,[203] reproducibility[204] and predictability.[205]

Limitations – A right to transparency and explainability should be made subject to limitations pursuant to the aforementioned three-part test, which could include among its legitimate aims intellectual property concerns, as well as considerations relating to the need to ensure the proper functioning of the AI system in question (in some contexts, too much information about the inner working of algorithms would enable to 'game the system' and manipulate their outcomes).[206] In addition, costs associated with disclosure

[202] See e.g., Shrutika Poyrekar, Traceability is how we lead AI, not just regulate it, *Medium*, 13 July 2025, https://medium.com/@mumbaiyachori/traceability-is-how-we-lead-ai-not-just-regulate-it-5461035adf3d.

[203] See e.g., Adrian Erasmus, Taylor D.P. Brunet, and Eyal Fisher, What is Interpretability?, 34 *Philos. & Technol.* (2020) 833.

[204] See e.g., Benjamin Haibe-Kains *et al*, Transparency and reproducibility in artificial intelligence, 586 *Nature* (2020) E14.

[205] See e.g., Aviad Raz *et al,* Prediction and explainability in AI: Striking a new balance? 11 *Big Data & Society* (2024), https://journals.sagepub.com/doi/10.1177/20539517241235871.

[206] See e.g., Stephan Grimmelikhuijsen, Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making, 83 *Public Adminstration Review* (2022) 241.

obligations should be factored in when assessing any disclosure requirements imposed on relevant duty holders (which may include, *inter alia*, developers and deployers of AI systems) in ways that resemble the imposition of conditions in connection with freedom of information requests in the offline world.[207]

A critical factor in implementing the right to explainability is the technological capacity to provide explanations that would be both intelligible and authentic. While interest in explainable AI (XAI) has been growing,[208] at this point it time, good levels of explainability may simply be unattainable. In such contexts, a choice may present itself between avoiding using sophisticated AI systems altogether and using them without being able to adequately explain their precise method of operation. Questions regarding the utility of the system and the harm if might inflict would inform any balancing test conduction in this connection.

5. The right not to be subject to algorithmic manipulation

Rationale – A right not to be subject to algorithmic manipulation should be developed in response to the exceptionally high potential of using AI systems to manipulate and steer choices and preferences in ways that threaten personal agency and human dignity. While problems related to manipulation of will and opinion are certainly not novel in nature, by collecting vast quantities of data about individual and collective decision-making and the factors that shape them, AI systems are particularly well-situated to allow the deployers of AI systems to influence individual and group decisions in much stronger ways than ever before. This is facilitated in part by the use of hyper-personalization or 'micro-targeting'[209] of individuals targeted by AI systems,[210] including a mapping of their preferences and emotional reactions, the ability to identify and exploit

[207] See e.g., The Freedom of Information Act. 5 USC Sec. 552 (USA).

[208] See e.g., Upol Ehsan and Mark O. Riedl, Social construction of XAI: Do we need one definition to rule them all?, 5 *Patterns* (2024), https://www.sciencedirect.com/science/article/pii/S2666389924000175.

[209] See e.g., Almog Simchon, Matthew Edwards and Stephan Lewandowsky, The persuasive effects of political microtargeting in the age of generative artificial intelligence, 3 *PNAS Nexus* (2024) 35.

[210] See e.g., Mrinalini Choudhary The Algorithmic Persuader: Ethical Challenges in AI-Powered Behavioral Manipulation in Digital Marketing, *RAIS Conference Proceedings* (2025), https://rais.education/wp-content/uploads/2025/05/0495.pdf.

cognitive biases and personal vulnerabilities, the deployment of subliminal forms of messaging and practices of constant and ubiquitous surveillance.

Pre-existing human rights norms protect some aspects of one’s cognitive “inner sanctum”[211] through the rights to freedom of thought and opinion,[212] and the right to privacy and human treatment.[213] In certain contexts, manipulation concerning the integrity of elections and data integrity could also run afoul of human rights norms concerning the right to vote and be elected [214] and the freedom to seek and receive information.[215] The interest in preserving data integrity might be particularly relevant in human rights regimes that recognize a right to the truth.[216] And, in addition, exploitation of pre-existing vulnerabilities (e.g., high rates of digital illiteracy among the elderly) may conflict with applicable non-discrimination norms.[217] Still, there is no comprehensive right against manipulation in the offline context, which broadly protects cognitive liberty and data integrity. A new right against AI manipulation could therefore serve as a response to the new threat landscape associated with AI technology.

[211] See e.g., Tim Bayne, *Thought: A Very Short Introduction* (2013) 34.

[212] ICCPR, art. 18, 19. See, for example, Joint Declaration on AI, Freedom of Expression and Media Freedom of 25 Oct. 2025 by the United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression, and the African Commission on Human and Peoples’ Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information in Africa (“mandate holders”), pp. 2-3 (“However, AI-driven personalisation and “micro-targeting” can manipulate opinion through sophisticated and non-consensual means interfering with individual agency. Gen AI can lead to personalised and interactive persuasion, leveraging individual behaviours and habits to steer exposure to certain information over time. Such practices can sway political perceptions, interfere with public discourse across borders, suppress dissenting views or favour government 2 narratives, radicalise individuals, or even undermine creativity and cognitive processes such as attention and critical thinking. These technologies can also pose specific risks for children, elderly, and vulnerable groups, harming mental and emotional health and development. Where AI undermines human autonomy or creativity, it unacceptably intrudes into individuals’ absolute right under international law to form their opinions free from non-consensual interference and manipulation. The opacity of AI tools aggravates the risk to freedom of opinion”).

[213] ICCPR, art. 7, 17.

[214] ICCPR, art. 25(b).

[215] ICCPR, art. 19(2).

[216] See e.g., *Lund v. Brazil,* judgment of the I/A Court of HR of 24 Nov. 2010, para. 211.

[217] See e.g., Monika Mayrhofer, Margit Ammer and Katrin Wladasch, The concept of vulnerability and its relation to equality in the context of human rights: cases from climate change, anti-discrimination and asylum, *Frontiers in Sociology* (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC11852298/pdf/fsoc-10-1522402.pdf.

Scope – The right not to be subject to algorithmic manipulation should capture a large set of influence techniques that are aimed at weakening or circumventing rational decision making and deliberative processes, through purposefully deceiving individuals or groups of individuals, or exploiting their cognitive biases and vulnerabilities. These techniques often fall below the assumption of full control over the thoughts and opinions of individuals (which freedom of thought and opinion already prohibit in absolute terms), and involve weaker forms of unacceptable influence that strive to steer minds and cognitive processes without absolutely controlling them. Such influence techniques may involve abuse of trust through human impersonation and other forms of misrepresentation, the dissemination of misinformation and disinformation, and the use of 'dark patterns'[218] and subliminal messaging. Associated practices such as constant and ubiquitous surveillance, emotion recognition and brain activity-tracking supporting efforts to steer individual or group conduct and govern choices (e.g., as part of social scoring and social conditioning programs)[219] could also come within the purview of the right.

Limitations – The right not to be subject to algorithmic manipulation should be regarded as a relative right by virtue of the challenging distinction between legitimate and illegitimate forms of manipulation.[220] Evaluation of the legitimacy of different manipulation practices should consider the means used, as well as the harms caused. Some forms of manipulation are subtle (e.g., social "nudges"),[221] some others involves trivial harms (e.g., advertisement aimed at steering choice between comparable products or services, or at purchasing cheap products or services), and yet others are aimed at advancing socially desirable changes in conduct or preference (e.g., reduce smoking or improve driving). There are also situations in which individuals consent to be

[218] See e.g., Tim Kollmer and Andreas Eckhardt, *Dark Patterns*, 65 *Business & Information Systems Engineering* (2023) 201.

[219] See e.g., Tuba Bircan and Mustafa F. Özbilgin, Unmasking inequalities of the code: Disentangling the nexus of AI and inequality, 211 *Technological Forecasting and Social Change* (2025), https://www.sciencedirect.com/science/article/abs/pii/S0040162524007236.

[220] See e.g., Sunstein, *Manipulation*, 37 *et seq*.

[221] See e.g., Richard A. Thaler and Cass A. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (2009) 74 *et seq*.

manipulated (e.g., engage with an AI companion as if it were a real person) or derive an objective or subjective benefit from being manipulated (e.g., in the context of virtual reality games). The development of a right not to be subject to AI manipulation might have to gradually develop use cases that give effect to the distinction between legitimate and illegitimate forms of manipulation (in ways comparable to distinction features in other rights – e.g., legitimate v. illegitimate distinctions in non-discrimination contexts). Increasing transparency – for example, by labelling artificial contents – and reducing harm – for instance, by subjecting certain uses of AI systems to close review and oversight – could also be part of such a balancing test.

In addition, the extent of the positive obligations associated with a right not to be manipulated would have to take into account the different costs, including human rights costs, of taking sweeping anti-manipulation measures such as filtering away all expressions of disinformation or misinformation. Measures like that might qualify as unacceptable forms of censorship and/or entail prohibitive costs on content intermediaries like online platforms.

6. The right to a human decision and a human-to-human interaction

Rationale – A right to a human decision and a human-to-human interaction may be developed in response to concerns about the arbitrary and non-transparent nature of algorithmic decisions and 'conduct'. It may also be viewed as a response to concerns about algorithmic safety, reliability and effectiveness, to concerns about algorithmic bias and unfairness, and to concerns about lack of algorithmic accountability. These concerns, at least partly, revolve around the compatibility of AI systems with human rights standards. This could justify, in and of itself, affording individuals an opt out option – that is, allowing individuals to choose not to interact with AI systems and be subject to their decision making authority in situations that threaten their human rights.

At a deeper level, developing a right to human decisions and interactions would reflect an anxiety about the complete substitution of human interlocutors with non-human ones, and its effects on societal and individual wellbeing. Arguably, the inability of AI systems to feel and authentically convey emotions, including empathy and compassion, when interacting with humans removes an important source of wellbeing from our social fabric when compared to interactions involving humans. What's more, the data-centric perspective through which algorithms relate to human beings has significant human dignity implications. It reduces humans in their 'eyes' from unique moral agents endowed with rich and multifaceted characteristics, feelings and aspirations, to a technical set of data in a manner that can be regarded as dehumanizing. It may also be regarded as humiliating for humans to be subject to the authority of machines and to be effectively excluded from meaningfully participating in decisions about their own lives.

Existing human rights norms deal only indirectly with the right to a human decision, and rarely, if ever, deal with the right to a human-to-human interaction. The aforementioned concerns about the quality of AI systems could translate themselves, if proven true, to human rights issues relating to decision-making in sectors in which such systems are deployed (health, education, employment, security of persons etc.), and concerns about algorithmic bias in algorithmic decisions and interaction could be tied to anti-discrimination norms. In the same vein, concerns about transparency and accountability regarding automated decisions could be connected to human rights prohibiting arbitrariness, and requiring reason-giving and an effective remedy. Still, concerns related to the inhumanity of AI systems are not well covered in existing human rights law, with the right not to be subject to inhuman treatment (developed in connection with the prohibition against torture)[222] not having been understood until now as a suitable vehicle for developing a right to be treated by humans.

Scope – The right to human decision should encompass decisions with significant implications, including decisions impacting legal rights. These may include consequential or high risk decisions regarding immigration status, eligibility for welfare benefits, job applications and job promotion, issuance of credit and provision of health treatment Decisions that relate to relatively trivial or non-risky matters, such as the content of social media feeds or eligibility for special promotion sales, should not be

[222] ICCPR, art. 7.

covered by the right. The protection afforded to decisions should extend to recommendations that are relied upon, in practice, as if they were decisions, or which are likely to be routinely followed.[223] In the same vein, interactions covered by a right to human-to-human interaction should include situations of a sensitive nature which have the potential for generating feelings of powerlessness, humiliation and anxiety, as well as other situations that could impact important needs and interests. These may include, for example, interactions between teacher-student, physician-patient, judge-defendant and welfare officer-welfare recipient.

The right to a human decision and human-to-human interaction should not be read as requiring the total exclusion of AI systems from relevant decision making processes and interactions. EU legislation,[224] which has strongly influenced the development of a right to a human decision in a number of other standard-setting instruments, focuses on the impermissibility of certain decisions based *solely* on automated systems, implying that human-machine interfaces that involves some division of labour between the humans and machines would not run afoul of the prohibition. Arguably, modalities that involve effective human oversight or an accessible and effective right to appeal automated decisions before humans could also be regarded as compatible with the right, provided that humans can exercise meaningful review and reconsideration in relation to the AI decision. In a similar manner, interactions involving both human and AI service providers would not generally be regarding as incompatible with the right to human-to-human interaction.

In all cases, the application of AI systems in non-trivial or non-obvious circumstances should be subject to the aforementioned notification and transparency requirements. This would allow individuals and groups of individuals to know about the involvement of AI systems in the decision or interaction, and be able to effectively enforce their right to a human decision and human-to-human interaction.

[223] Case C-634/21, OQ v. Land Hessen (Schufa case), CJEU judgment of 7 Dec. 2023, ECLI:EU:C:2023:957.
[224] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 23.11.1995, OJ L 281/31; GDPR, art. 22.

Limitations – Like other AI human rights, the right to a human decision and human-to-human interaction is relative in nature, and should be made subject to the three-part test mentioned above (legality, legitimacy and necessity and proportionality). There can be contexts where scale, cost and efficiency considerations would militate in favour of reliance on automated systems with regard to matters in relation to which reasonable expectations of automation can be found (or emerge over time) – e.g., when calculating tax rates or issuing parking tickets. The acceptability of such automated decisions might depend on whether or not effective appeal procedures are in place to challenge inaccurate decisions (such procedures may entail the imposition of reasonable costs and procedural requirements on those challenging the automated decision). Similar considerations pertaining to reasonable expectations might also govern the evaluation of measures seeking to replace human-to-human interactions with machine-centred interactions (e.g., for vocational training purposes or for handling routine consumer complaints).

In all cases covered by the right, however, the normative point of departure should be respecting – as much as possible – individuals' choice to opt out from engaging with AI systems or to challenge their decisions before humans. From this perspective, the exceptions found in article 22 of the GDPR – necessity for contract performance, authorization by a law that introduces alternative suitable measures to protect the rights and interests of the affected "data subject" and explicit consent – appear to be too generous in nature, and not fully compatible with an understanding of the GDPR right to a human decision as a full-fledged human right.

With regard to consent, there should be circumstances when consent would not be sufficient to allow for exclusive reliance on algorithms, given the societal importance of maintaining close human involvement in certain decisions and interactions. This is the case, for example, with regard to decision-making in criminal cases, and in life and death decisions taken in the healthcare system. In addition, the ability to manipulate consent through the deployment of AI systems supports consigning only limited weight to expressions of consent in situations where AI decisions or interactions could cause serious harm to individuals. Still, if appropriate safeguards are in place, the choice of individuals who show a clear preference towards AI decisions and interactions should be respected when possible.

7. The right to accountability for harms caused by the use of AI systems

Rationale – A right to accountability for the harms caused by the use of AI systems is required in order to effectively protect other human rights adversely affected by the use or failure to use AI systems. This is the AI equivalent to the right to an effective remedy that already exists in international human right law and which constitutes a critical safeguard for maintaining the effectiveness and fairness of human rights regimes.[225] Specifically, article 2(3)(a) of the ICCPR requires states to ensure that "any person whose rights or freedoms as herein recognized are violated shall have an effective remedy, notwithstanding that the violation has been committed by persons acting in an official capacity" and sub-paragraphs (3)(b) and (3)(c) require states to provide judicial or other procedures for determining violations and to enforce remedies. The need for accountability is particularly acute with regard to AI systems, given the "many hands" problem[226] associated with attaching responsibility for harms their use entails (given the involvement of a long supply chain, including designers, developers, disseminators, users, overseers etc.), the limited ability to foresee such harms, the challenge of establishing intent in connection with acts and omissions meditated by sophisticated AI systems, and the general problem of assigning human rights responsibility to business entities and other non-state actors (which is the opposite of the problem of "official capacity" alluded to in article 2(3)(a)).

Developing a right to accountability in this context would build upon other legal initiative in this domain, such as the UN Guiding Principles on Business and Human Rights, the EU Corporate Sustainability Due Diligence Directive and the provisions of

[225] See e.g., ICCPR, art. 2(3).

[226] See e.g., Helen Nissenbaum, Accountability in a computerized society, 2 *Science and Engineering Ethics* (1996) 25.

the African Charter.[227] It would complement such instruments by addressing some of the specific challenges posed by the use of AI systems and clarify the division of labour between states and non-state actors, including international organizations, relating to assigning responsibility for remediation and operationalizing the right to a remedy in concrete use cases.

Scope – A right to accountability should correspond to all stages in the life cycle of AI systems, encompassing the responsibility of all individuals and companies, states and international organizations involved in the design, development, dissemination, use, review and oversight of such systems. Accountability in this realm should involve elements of risk and impact assessment, risk and impact mitigation, pre-launch testing, ongoing review and oversight, handling of complaints and addressing new information on risk and impact. Accountability regimes should require the introduction of both pro-active and reactive safeguard measures and the assignment of legal responsibility for failing to properly carry them out. Given the systemic nature of most prophylactic and reparative measures, they should not be necessarily tied to claims by any specific individual, but rather be viewed as part of the implementation of the broad obligation of states to ensure human rights, including by managing the technological ecosystem, and of the responsibility of business entities to respect human rights and to exercise due diligence across their value chain.[228]

Yet, accountability mechanisms should also be able to assign, in appropriate cases, direct responsibility for harms caused by different actors involved in developing and using any specific AI system. Such measures of accountability inevitably depend on the degree of transparency afforded with relation to the system in question, including with regard to allocation of tasks and functions across the value chain. They also depend on adapting liability laws, such as tort law, contract law and criminal law, to the unique challenges posed by algorithmic unpredictability and non-traceability

[227] See African Charter, art. 27 and 28.
[228] See e.g., UNGPs, p. 19.

through the imposition of measures like strict civil liability or the reversal of burdens of proof in certain cases.

The right to accountability vis-à-vis uses of AI systems should follow the logic of article 2(3) of the ICCPR, confirming the right to remedy of those harmed by the use of such systems – including identifying relevant duty holders who would be expected to respect and ensure the right – the right of victims to obtain a determination of the violation of their rights through judicial or non-judicial avenues, and their right to enforce the remedy due to them. Given the multi-jurisdictional dimensions of harms caused by the use of AI systems, states should develop ways to cooperate with one another and with international organizations and business entities with a view to implementing across different jurisdictions the right to accountability.

Limitations – The right to accountability is limited in nature, due to the many restrictions legal systems impose on assigning legal responsibility through standards of proof, establishment of intent or negligence, problems of causation, jurisdiction etc. Furthermore, the development of AI systems often relies on "safe harbours",[229] which link implementation of best industry standards for harm prevention and mitigation with liability shields. Such schemes could be justified on the basis of the need to promote innovation and sustainable technological development, but they ought to be set off against general compensation funds or comparable remedial measures, as well as engaging in lesson learning and acknowledging 'moral responsibility'. In cases involving the imposition of direct or indirect liability on entities involved in the development and deployment of AI systems, the extent of any reparations issued should be proportionate to the degree of harm predictability and to any due diligence efforts undertaken by relevant actors to prevent or mitigate the harm caused.

[229] See e.g., Nicoletta V. Kolpakov, AI's Escalating Sophistication Presents New Legal Dilemmas, *NYSBA,* 28 May 2025, https://nysba.org/ais-escalating-sophistication-presents-new-legal-dilemmas/?srsltid=AfmBOoqXwvLfl5UBc19zMUaBSYn88kci52f6zqw4Hm6h1f58E24GcUqE.

Conclusions:

The use of AI systems raises many significant human rights issues. Some of them relate to the exceptional power and effectiveness of AI systems which could be harnessed for enhancing right enjoyment, but also for eroding existing human rights. AI systems may also be used to circumvent existing human rights safeguards and to undermine existing systems of accountability for human rights violations. In addition, unique features of AI systems – in particular, the ‘black box’ and their lack of humanity – create unique problems of non-transparency and dehumanization – which arguably merits a specific human rights response.

Existing international human rights law standards have much to offer in connection with the use of AI systems: The right to privacy, non-discrimination, the right to seek and receive information, and the right to enjoy the benefits of scientific progress are relevant here, alongside many other human rights. Still, these classic human rights afford only limited guidance in respect of the unique challenges posed by AI systems and the modalities for their development and deployment. They also fail to effectively protect some important aspects of the use or lack of use of AI systems. As in other areas of human rights law, a process of normative elaboration and gap-filling could be useful, and having democratically-representative bodies partake in them would increase the legitimacy of the standards emerging out of this process.

A review of recent standard-setting instruments corresponding, at some level, to the need and interest in protecting individual rights against certain uses of AI systems, suggests that these standards do not offer a clear, comprehensive and effective human rights response to the relevant challenges. They are often formulated in language that differs from that of international human rights instruments, they tend to be too abstract or too specific, and they usually address AI systems in a piecemeal manner – sometimes dealing with one AI human right only. They also tend to be non-binding, offer generous exceptions and do not embed themselves in human rights systems of protection.

It is against this background that this White Paper – building on extensive consultations – proposes to move forward with the articulation of a draft soft law instrument that could serve as blueprint for a future international AI bill of human rights. The choice of soft law is dictated both by practical considerations relating to what can be realistically achieved at this moment in time, but also from principled considerations regarding to the need to engage with a broad range of democratically-elected and deliberative bodies in the actual formulation of any instrument that would ultimately have a legally binding effect. Still, even an initial proposal could help interpretative bodies involved to give greater effect to human rights in AI contexts and it might inspire further discussions, consultations and democratic deliberations on these issues.

The White Paper offers an initial list of seven rights, which I recommend for inclusion in a future AI bill of human rights. These rights seek to protect our very ability to use AI systems in ways compatible with human wellbeing, in ways that minimize harm to humans and which meets standards of fairness and justice in the applicable processes and outcomes attendant to the development of use of such systems:
The right of access to AI systems
The right to privacy-related protections from harmful uses of AI systems
The right to be free from algorithmic bias and unfairness
The right to algorithmic transparency and explainability
The right not to be subject to algorithmic manipulation
The right to a human decision and a human-to-human interaction
The right to accountability for harms caused by the use of AI systems

The list is built on my interpretation of the literature on AI and human rights, existing international human rights law norms, the reviewed standard setting instruments and the consultations I undertook. Hopefully, it can serve as a basis for future normative work in the field by different stakeholders and constituencies. Given the breakneck speed of developments in AI technology, there is little time to waste in developing an effective human rights response. This White Paper strives to help to accelerate such a response