Research Paper

International Law Programme

January 2023

AI governance and human rights

Resetting the relationship

Kate Jones

CHATHAM HOUSE

Chatham House, the Royal Institute of International Affairs, is a world-leading policy institute based in London. Our mission is to help governments and societies build a sustainably secure, prosperous and just world.

Contents

Summary

- Artificial intelligence (AI) is redefining what it means to be human. Human rights have so far been largely overlooked in the governance of AI – particularly in the UK and the US. This is an error and requires urgent correction.
- While human rights do not hold all the answers, they ought to be the baseline for AI governance. International human rights law is a crystallization of ethical principles into norms, their meanings and implications well-developed over the last 70 years. These norms command high international consensus, are relatively clear, and can be developed to account for new situations. They offer a well-calibrated method of balancing the rights of the individual against competing rights and interests using tests of necessity and proportionality. Human rights provide processes of governance for business and governments, and an ecosystem for provision of remedy for breaches.
- The omission of human rights has arisen in part because those with human rights expertise are often not included in AI governance, both in companies and in governments. Various myths about human rights have also contributed to their being overlooked: human rights are wrongly perceived as adding little to ethics; as preventing innovation; as being overly complex, vague, old-fashioned or radical; or as only concerning governments.
- Companies, governments and civil society are retreading the territory of human rights with a new proliferation of AI ethics principles and compliance assessment methods. As a result, businesses developing or purchasing AI do not know what standards they should meet, and may find it difficult to justify the costs of ethical processes when competitors have no obligation to do the same. Meanwhile, individuals do not know what standards they can expect from AI affecting them and often have no means of complaint. Consequently, many people do not trust AI: they suspect that it may be biased or unfair, that it could be spying on them or manipulating their choices.
- The human rights to privacy and data protection, equality and non-discrimination are key to the governance of AI, as are human rights' protection of autonomy and of economic, social and cultural rights in ensuring that AI will benefit everyone. Human rights law imposes not only duties on governments to uphold, but also responsibilities on companies and organizations to comply, as well as requirements for legal remedies and reparation of harms.
- Companies and investors, governments, international organizations and civil society should take steps to establish human rights as the foundation on which AI governance is built, including through inclusive discussion, championing human rights and establishing standards and processes for implementation of human rights law and remedy in case of breach.

01 Introduction

AI is redefining what it means to be human. As existing international norms designed to allow every human being a life of liberty and dignity, human rights ought to be the foundation for AI governance.

Human rights are central to what it means to be human. They were drafted and agreed internationally, with worldwide popular support, to define freedoms and entitlements that would allow every human being to live a life of liberty and dignity. Those fundamental human rights have been interpreted and developed over decades to delineate the parameters of fairness, equality and liberty for every individual.

Now, artificial intelligence (AI) is redefining what it means to be human. Its systems and processes have the potential to alter the human experience fundamentally. AI will affect not only public policy areas such as road safety and healthcare, but also human autonomy, relationships and dignity. It will affect lifestyles and professions, as well as the future course of human development and the nature and scale of conflicts. It will change the relationships between communities and those between the individual, the state and corporations.

AI offers tremendous benefits for all societies but also presents risks. These risks potentially include further division between the privileged and the unprivileged; the erosion of individual freedoms through ubiquitous surveillance; and the replacement of independent thought and judgement with automated control.

This paper aims to explain why human rights ought to be the foundation for AI governance, to explore the reasons why they are not – except in the EU and some international organizations – and to demonstrate how human rights can be embedded from the beginning in future AI governance initiatives.

While AI is being implemented rapidly around the world, most governance initiatives to date have emerged from developed states. This paper therefore focuses on practice and process primarily in the EU, the UK and the US. However, the paper also acknowledges the significance of AI initiatives elsewhere in the world – China in particular is a leading developer and exporter of AI technology.

The following chapter explains AI and the risks and benefits it presents for human rights. Chapter 3 aims to dispel myths and fears about human rights, before discussing why human rights should provide the baseline for AI governance. Chapters 4, 5 and 6 outline the principal import of human rights for AI governance principles, processes and remedies respectively. Finally, Chapter 7 offers recommendations on actions that governments, organizations, companies and individuals can take to ensure that human rights are embedded in AI governance in future.

02
What is AI?

AI has capacity to transform human life – both for better and for worse.

AI is increasingly present in our lives, and its impact will expand significantly in the coming years. From predictive text, to social media news feeds, to virtual homes and mobile phone voice assistants, AI is already a part of everyday life. AI offers automated translation, assists shoppers buying online and recommends the fastest route on the drive home. It is also a key component of much-debated, rapidly developing technologies such as facial recognition and self-driving vehicles.

There is no single agreed definition of AI: it is a general term referring to machines' evolving capacity to take on tasks requiring some form of intelligence. The tasks that AI performs can include generating predictions, making decisions and providing recommendations.[1] This means that AI may make decisions itself, or provide information for use in human decision-making.

AI systems are algorithmic – the algorithm being the computational process or set of rules that the computer follows to calculate a result. To learn, AI generally relies on synthesising and making inferences from large quantities of data. It is the machine's capacity to learn by itself how to do tasks better, rather than simply following instructions, that distinguishes AI from traditional computer programmes. Contrary to popular myth, self-improvement does not prevent AI from being constrained by rules.

Governments are among the largest adopters of AI, deploying it to assist in making decisions that can have major consequences for the lives of individual citizens. For example, governments are using AI to assist with decisions on entitlement to immigration status, welfare benefits, school entry and priority

1 The European Commission's High-Level Expert Group on Artificial Intelligence offers a fuller definition: 'Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.' Independent High-Level Expert Group on Artificial Intelligence (2019), *Ethics Guidelines for Trustworthy AI*, Brussels: European Commission, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

vaccinations. They are adopting it to assist with provision of justice, in both civil and criminal processes. And they may be using AI to assist in delivery of critical infrastructure and national security.

AI is likely to pervade almost every domain of human activity, and to become increasingly important as technology evolves towards greater interoperability, including through the development of the metaverse.[2] This paper discusses general features of AI, but by no means diminishes the need for parallel sector-specific discussion. The use of AI in the healthcare system, in social media or in the criminal justice process, for instance, each raise specific human rights issues that need to be addressed in context, alongside the overarching issues discussed here.

2.1 What potential does AI hold for human rights and the common good?

Due to its speed and its power of self-learning, AI has the capacity to transform our societies. It can operate faster – and potentially better – than any human. It can achieve scientific breakthroughs, calculate fair distributions and outcomes, and make more accurate predictions.

AI holds enormous potential to enable human development and flourishing. For example, AI is accelerating the battle against disease[3] and mitigating the impact of disability;[4] it is helping to tackle climate change[5] and optimize efficiency in agriculture;[6] it can assist distribution of humanitarian aid;[7] it has enormous potential for improving access to, and quality of, education globally;[8] and it can transform public and private transport.[9] AI could help to ensure that policing is fair and respectful of human dignity. It may make workplaces more productive, reduce the load of manual labour and help developed countries to manage the challenges of an ageing population. To give a specific example of the benefits, the AI programme AlphaFold is predicting the structures of both human and animal proteins with tremendous speed and remarkable accuracy, with potentially transformative effects on medical treatments, crop science and plastic reduction.[10]

---

2 Moynihan, H., Buchser, M. and Wallace, J. (2022), *What is the metaverse?*, Explainer, London: Royal Institute of International Affairs, https://www.chathamhouse.org/2022/04/what-metaverse.
3 For example, as regards COVID-19: Soomro, T. A. et al. (2021), 'Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): a detailed review with direction for future research', *Artificial Intelligence Review*, 55(2), pp. 1409–39, https://doi.org/10.1007%2Fs10462-021-09985-z.
4 For example: Microsoft (undated), 'AI for Accessibility', https://www.microsoft.com/en-us/ai/ai-for-accessibility.
5 Rolnick, D. et al. (2019), 'Tackling Climate Change with Machine Learning', *arXiv*, 1906.05433v2 [cs.CY], https://arxiv.org/pdf/1906.05433.pdf.
6 Cline, T. (2019), 'Digital agriculture: making the most of machine learning on farm', Spore, https://spore.cta.int/en/dossiers/article/digital-agriculture-making-the-most-of-machine-learning-on-farm-sid0dbfbb123-30b2-48fd-830e-71312f66af04?msclkid=d9322204a57311ecb7a36f2895e35dd1.
7 For example: UN Global Pulse (2022), 'Innovating Together for our Common Future', www.unglobalpulse.org.
8 For example: UNESCO (2022), 'Artificial Intelligence and the Futures of Learning', https://en.unesco.org/themes/ict-education/ai-futures-learning.
9 For example, European Parliament Briefing (2019), 'Artificial Intelligence in Transport: Current and Future Developments, Opportunities and Challenges', https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRS_BRI(2019)635609_EN.pdf?msclkid=cd1a70d2aaa011ec9f9ff79af4f9d88d.
10 Tunyasuvunakool, K. et al. (2021), 'Highly accurate protein structure prediction for the human proteome', *Nature*, 596, 21 July 2021, pp. 590–96, https://doi.org/10.1038/s41586-021-03828-1.

In short, when properly managed, AI can enable delivery of the UN's Sustainable Development Goals (SDGs) by the 2030 deadline,[11] boost the implementation of economic, social and cultural rights worldwide, and support improvements in many areas of life.

To achieve these aims, AI must be harnessed for the good of all societies. Doing so involves not only goodwill, but also ensuring that commercial considerations do not dictate the development of AI entirely. Provision of funding for AI research and development outside the commercial sector will be invaluable, as will access to data for AI developers such that they may generate applications of AI that benefit people in all communities.

Just as the industrial revolution brought progress at the expense of upheaval in traditional ways of living, so will AI bring change to our societies. Work must be done now to mitigate the risk of negative impacts. Governments must anticipate and manage the changes that widespread use of AI will herald. They must consider both the implications of AI for their own public policymaking, which may be subject to judicial review, and how to govern a society in which AI is increasingly being developed by the private sector and becoming a feature of life for the world's population. This includes governance not only of AI itself but of its implications for current ways of life. For example, governments should address the risk that AI will upend current practices and norms in the workplace, through mass unemployment and an undermining of bargaining power between employers and employees. Governments should be taking active steps to ensure the benefits of AI are distributed equitably, avoiding the division of society into 'winners' and 'losers' from emerging technology. To preserve and promote public interest, governments must not allow companies to develop AI in a policy and regulatory vacuum.

2.2 What are the key human rights and ethical challenges posed by AI?

Evidence abounds of problematic uses of AI. At one end of the spectrum, AI is being deliberately used as a tool of suppression: for example, the Chinese government's use of AI to conduct mass surveillance of its Uyghur minority.[12] Some types of AI could be used deliberately to limit people's freedom to express themselves and to meet with others, to monitor the general public for compliance with behavioural rules,[13] to detect 'suspicious behaviour'[14] or to restrict access to society's benefits to a privileged few.

---

11 Vinuesa, R. et al. (2020), 'The role of artificial intelligence in achieving the Sustainable Development Goals', *Nature Communications*, 11(233), https://doi.org/10.1038/s41467-019-14108-y; Chui, M. et al. (2019), 'Using AI to help achieve Sustainable Development Goals', New York: UN Development Programme, https://www.undp.org/blog/using-ai-help-achieve-sustainable-development-goals.
12 Mozur, P. (2019), 'One Month, 500,000 Face Scans: How China is using AI to profile a minority', *New York Times*, 14 April 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.
13 Heikkila, M. (2021), 'The rise of AI surveillance', Politico, 26 May 2021, https://www.politico.eu/article/the-rise-of-ai-surveillance-coronavirus-data-collection-tracking-facial-recognition-monitoring.
14 Vinocur, N. (2020), 'French politicians urge deployment of surveillance technology after series of attacks', Politico, 30 October 2020, https://www.politico.eu/article/french-politicians-urge-deployment-of-surveillance-technology-after-series-of-attacks.

Many AI tools abuse human rights as a collateral consequence of their operation. AI risks embedding and exaggerating bias and discrimination, invading privacy, reducing personal autonomy and making society more, rather than less, unequal. For example, AI sentencing tools may discriminate against minorities, potentially turning back decades of progress towards equality. AI in healthcare may harm human health if algorithms are incorrect or biased,[15] while AI in welfare-provision or migration may make unfair decisions on eligibility. AI tools may infer sensitive information about individuals in violation of their privacy.

Even an AI tool designed with the intention of implementing scrupulous standards of fairness will fail if it does not replicate the complex range of factors and subtle, context-specific decision-making processes of humans. Unchecked, AI systems tend to exacerbate structural imbalances of power and to disadvantage the most marginalized in society.[16]

Further, some AI tools may have outputs detrimental to humanity through their potential to shape human experience of the world. For example, AI algorithms in social media may, by distorting the availability of information, manipulate audience views in violation of the rights to freedom of thought and opinion,[17] or prioritize content that incites hatred and violence between social groups.[18] AI used to detect aptitudes or to select people for jobs, while intended to broaden human horizons and ambition, risks doing the opposite. Without safeguards, AI is likely to entrench and exaggerate social divides and divisions, distort our impressions of the world and thus have negative consequences on aspects of human life. These risks are amplified by the difficulty of identifying when AI fails, for example when it is malfunctioning, manipulative, acting illegally or making unfair decisions. At present, companies rarely make public their identification of mistakes or errors in their AI. Consumers cannot therefore see which standards have been met.

Finally, AI may entrench and even exacerbate social divides between rich and poor, worsening the situation of the most vulnerable. As AI development and implementation is largely driven by the commercial sector, it risks being harnessed for the benefit of those who can pay rather than to resolve the world's most significant challenges, and risks being deployed in ways that further dispossess vulnerable communities around the world.[19]

---

15 Park, Y. et al. (2020), 'Evaluating artificial intelligence in medicine: phases of clinical research', *JAMIA Open*, 3(3), October 2020, pp. 326–31, https://doi.org/10.1093/jamiaopen/ooaa033.
16 European Digital Rights (EDRi) et al. (2021), *Civil Society Statement on an EU Artificial Intelligence Act for Fundamental Rights*, 30 November 2021, https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf; Kalluri, P. (2020), 'Don't ask if artificial intelligence is good or fair, ask how it shifts power', *Nature*, 7 July 2020, https://www.nature.com/articles/d41586-020-02003-2.
17 Jones, K. (2019), *Online Disinformation and Political Discourse: Applying a Human Rights Framework,* Research Paper, London: Royal Institute of International Affairs, https://www.chathamhouse.org/2019/11/online-disinformation-and-political-discourse-applying-human-rights-framework.
18 Kornbluh, K. (2022), 'Disinformation, Radicalization, and Algorithmic Amplification: What Steps Can Congress Take?', Just Security blog, 7 February 2022, https://www.justsecurity.org/79995/disinformation-radicalization-and-algorithmic-amplification-what-steps-can-congress-take.
19 Hao, K. (2022), 'AI Colonialism', MIT Technology Review, 19 April 2022, https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism.

03
Governing AI: why human rights?

Human rights have been wrongly overlooked in AI governance discussions. They offer clarity and specificity, international acceptance and legitimacy, and mechanisms for implementation, oversight and accountability.

In the 1940s, there was fervent belief that human rights would be central to world peace and to human flourishing, key not only to safeguarding humanity from catastrophe but to the enjoyment of everyday life.[20] Supporters of the 'vast movement of public opinion'[21] in favour of human rights at that time would be amazed at their relative absence from today's debate on AI.

3.1 Human rights overlooked

AI governance has much to gain from a multidisciplinary (and potentially interdisciplinary) approach, drawing from, among others, philosophy, human rights law, science and technology studies, sociology, statistics, diverse impact assessment and audit practices and stakeholder theory. However, with some exceptions,[22]

20 See David Maxwell Fyfe's closing speech for the UK prosecution at Nuremberg, available at 'The Human's In the Telling', https://thehumansinthetelling.wordpress.com.
21 René Brunet, former delegate to the League of Nations, quoted in Lauren, P.G. (2011), *The Evolution of International Human Rights: Visions Seen*, 3rd edn, Philadelphia: University of Pennsylvania Press, p. 153.
22 Exceptions include the EU's Artificial Intelligence Act (discussed below) and academic texts including: McGregor, L., Murray, D. and Ng, V. (2019), 'International Human Rights Law as a Framework for Algorithmic Accountability', *International & Comparative Law Quarterly*, 68(2), April 2019, pp. 309–43, https://doi.org/10.1017/S0020589319000046; and Yeung, K., Howes, A. and Pogrebna, G. (2019), 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight: an end to Ethics Washing', in Dubber, M. and Pasquale, F. (eds) (2019), *The Oxford Handbook of AI Ethics*, Oxford: Oxford University Press. The White House's recent 'Blueprint for an AI Bill of Rights' helpfully introduces the language of rights into mainstream AI governance in the US, albeit without focusing directly on the existing international human rights framework. See The White House Office of Science and Technology Policy (2022), 'Blueprint for an AI Bill of Rights: Making Automated Systems Work For The American People', https://www.whitehouse.gov/ostp/ai-bill-of-rights.

the human rights framework has been overlooked as an existing and flexible baseline for AI governance.

AI governance initiatives are often branded as 'AI ethics', 'responsible AI' or 'value sensitive design'. Some of these initiatives, such as the Asilomar AI Principles,[23] are statements drawn primarily from the philosophical discipline of ethics. Many are multidisciplinary statements of principle, and so may include human rights law as an aspect of 'ethics'. For example, the UNESCO Recommendation on the Ethics of Artificial Intelligence lists '[r]espect, protection and promotion of human rights and fundamental freedoms and human dignity' as the first of its 'values' to be respected by all actors in the AI system life cycle.[24] And the Institute of Electrical and Electronics Engineers (IEEE)'s Standard Model Process for Addressing Ethical Concerns during System Design lists as its first 'ethical principle' that '[h]uman rights are to be protected'.[25]

Many sets of AI governance principles produced by companies, governments, civil society and international organizations fail to mention human rights at all.

Many sets of AI governance principles produced by companies, governments, civil society and international organizations fail to mention human rights at all. Of those that do, only a small proportion (around 15 per cent) take human rights as a framework.[26] Most national AI strategies do not engage with human rights in depth.[27]

So why, then, are human rights not central to AI governance?

First, in many arenas, human rights are simply omitted from discussions on AI governance. Software developers and others in the AI industry generally do not involve anyone from the human rights community in discussions on responsible AI. There is a marked lack of human rights-focused papers or panels at the largest

---

23 Future of Life Institute (2017), *Asilomar AI Principles*, https://futureoflife.org/2017/08/11/ai-principles.
24 UN Educational, Scientific and Cultural Organization (2021), *Recommendation on the Ethics of Artificial Intelligence*, Paris: UNESCO, https://unesdoc.unesco.org/ark:/48223/pf0000381137, section III.1.
25 Institute of Electrical and Electronics Engineers (IEEE), Standard Model Process for Addressing Ethical Concerns during System Design, IEEE Std 7000-2021, Annex H.
26 A 2020 review of 36 prominent sets of AI principles from around the world, authored by a diverse range of governmental and non-governmental bodies, found that only 23 referred to international human rights. Only one-half of the government documents reviewed include any reference to human rights. Five of the 36 sets of AI principles used international human rights as a framework for their work. See Fjeld, J. et al. (2020), *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,* Berkman Klein Center for Internet & Society, Research Publication No. 2020-1, http://dx.doi.org/10.2139/ssrn.3518482. A separate evaluation of 22 sets of guidelines makes no reference to human rights: Hagendorff, T. (2020), 'The Ethics of AI Ethics: An Evaluation of Guidelines', *Minds and Machines*, 30, pp. 99–120, https://link.springer.com/article/10.1007/s11023-020-09517-8.
27 A review of all national AI strategies published by 1 January 2020 found that while a majority referred to human rights, most only mentioned them in passing rather than engaging with them in depth. Only European states and India referred to human rights; many East and South East Asian states had developed a strategy, but these did not refer to human rights. None of the strategies reviewed came from Africa, and only one from Latin America (Colombia). Global Partners Digital and Stanford Global Digital Policy Incubator (2020), *National Artificial Intelligence Strategies and Human Rights: A Review*, https://www.gp-digital.org/wp-content/uploads/2020/04/National-Artifical-Intelligence-Strategies-and-Human-Rights%E2%80%94A-Review_April2020.pdf.

international conferences on responsible AI.[28] Corporate-level discussion of AI ethics and their implementation often fails to refer to, or engage with, human rights. Job advertisements for corporate AI ethics specialists usually make no reference to human rights. Governments focused on AI ethics may not involve human rights lawyers in policy development until a late stage, if at all. In contrast, human rights are often the focus of civil society and academic discussions in different venues – and with different participants – to those where corporate and public sector AI governance decisions are made.[29] Notable exceptions are discussions hosted by international organizations such as the UN and the Council of Europe, where human rights law forms a well-established shared lexicon; and the European Union, which has placed human rights at the core of the draft Artificial Intelligence Act.[30]

Second, certain myths about human rights can too often lead to them being disregarded by those involved in AI governance discussions. The following are some of the most common.

3.2 Myths about human rights

Myth 1. 'Ethics holds all the answers'

Ethics and human rights are distinct disciplines with valuable, complementary roles to play in AI governance. Both ethics and human rights share the rationale of curbing state and corporate power by acting as a bulwark of the interests of the individual. But they offer different, complementary means for reaching this end. One cannot substitute for the other or be considered at the exclusion of the other. Both disciplines must be considered together.

Ethics plays an important role in preceding and supplementing regulation. It has been the subject of much pioneering research and implementation in the field of AI governance. However, ethics is a branch of philosophy, not a system of norms: multiple versions are possible, and – despite, or perhaps exacerbated by, the efforts to draft so many sets of AI ethics principles – there is currently a lack of international consensus as to what precisely AI ethics entails. Significant differences of both substance and terminology between these sets of principles make it difficult for companies and public bodies to understand their responsibilities, and for individuals to know what standards to expect.

---

28 See the analysis of research contributions and shortcomings, including significant influence of industry, at the ACM Conference on Fairness, Accountability and Transparency: Laufer, B. et al. (2022), 'Four years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects', ACM Digital Library, https://doi.org/10.1145/3531146.3533107.
29 Of over 180 papers accepted for the ACM Conference on Fairness, Accountability and Transparency 2022 – 'a computer science conference with a cross-disciplinary focus' – only three refer to human rights in their abstract. ACM FAccT (2022), 'Accepted Papers', https://facctconference.org/2022/acceptedpapers.html. In contrast, Access Now's RightsCon Conference 2022, on technology and human rights, included Artificial Intelligence as one of its programme tracks; but only 11 per cent of its attendees came from the private sector and 4 per cent from government. RightsCon (2022), *Outcomes Report*, https://www.rightscon.org/cms/assets/uploads/2022/09/Outcomes-Report-2022-v10.pdf.
30 Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

The malleability of ethics means that it is difficult for civil society to hold other actors to account. Some technology companies face criticism for so-called 'ethics-washing' undertaken for reputational purposes,[31] and for exerting undue influence on some ethics researchers through funding.[32] Courts and tribunals do not allocate remedies for compliance with ethics. Moreover, while ethical principles are intended to ensure that technology reflects moral values, a focus on ethics may minimize the appetite for legal regulation.[33]

Although in some environments, the branding of 'ethics' may be more palatable than that of human rights for political reasons, it is of primary importance that human rights are considered at all – whatever the branding. To avoid conceptual confusion, human rights ought to be regarded as parallel to ethics rather than as a mere element of it. Any principles and processes of ethics should complement, rather than compete with, the existing human rights legal system. Conflicts between norms are damaging as they undermine the legal certainty and predictability of regulatory behaviour on which states, businesses and individuals rely.

Current popular support for AI ethics in principle, without a shared understanding of what AI ethics means in practice, has similarities with support for human rights in the 1940s. There was widespread support then for the concept of human rights, to prevent a repetition of the atrocities of the Second World War and to end domination and repression. However, there was no specific understanding or consensus on what exactly 'human rights' meant. Establishing agreement on the content of the Universal Declaration of Human Rights – and later the International Covenant on Civil and Political Rights and International Covenant on Economic, Social and Cultural Rights – required worldwide canvassing, expert input, negotiation and political compromise.[34] There is no evidence that reaching universal agreement on AI ethics without reference to the already-agreed human rights framework would be easier, or less politically charged, than those 20th century debates.

Myth 2. 'Human rights prevent innovation'

Human rights do not prevent innovation or undermine a 'move fast and break things' ethos, save that they entail compliance with minimum standards and therefore forbid certain egregious activities. Most innovators want a level playing field, and to avoid being undercut by actors with lower standards or being caught in a 'race to the bottom' with unscrupulous competitors. Innovators want to know how they can meet shared standards and inspire trust in their products. Human rights provide an appropriate basis for standards and processes internationally. For businesses, considering human rights from the outset of AI development and deployment may help to foster customer trust and minimize potential costs and time expended in litigation at a later stage.

---

31 Carnegie Council for Ethics in International Affairs (undated), 'Ethics Washing', https://www.carnegiecouncil.org/explore-engage/key-terms/ethics-washing (accessed 12 Sep. 2022).
32 Williams, O. (2019), 'How Big Tech funds the debate on AI ethics', New Statesman, 6 June 2019, https://www.newstatesman.com/science-tech/2019/06/how-big-tech-funds-debate-ai-ethics.
33 Wagner, B., (2018), 'Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping?' in Bayamlioglu, E. et al. (eds) (2018), *Being Profiled:Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, pp. 84–8, Amsterdam: Amsterdam University Press, https://doi.org/10.1515/9789048550180-016.
34 Lauren (2011), *The Evolution of International Human Rights*.

Myth 3. 'Human rights are complex and entail expensive legal advice'

While human rights can appear complex to non-specialists, initiatives such as the UN's B-Tech project show how the technology industry and investors can implement their human rights responsibilities.[35] Routine inclusion of human rights in computer science and coding training could reduce the perception of complexity. In reality, human rights are no more complex than any equivalent system of rules or principles: they consist of clear rules, with steps to be followed in implementing them. While novel situations will still pose challenges, human rights have been developed over many years and are inherently flexible to adapt to such challenges. In this way, human rights have answers for many situations, in terms of steps to follow or outcomes to reach.

A business trying to establish ethical credentials needs advice in order to do so effectively – this is the case whatever the source of the rules followed. Following human rights standards means following relatively clear, existing rules and minimizing the chances of public censure or litigation for failure to comply.

Myth 4. 'Human rights are about governments'

Human rights are not commonly part of the lexicon of AI developers and corporate ethics advisers – particularly outside the EU – because they are seen as regulating government, rather than corporate, activity.

While states are the primary bearer of duties under international human rights law, all companies have responsibilities to respect human rights. The Office of the UN High Commissioner on Human Rights (OHCHR)'s Guiding Principles on Business and Human Rights, unanimously endorsed by the UN Human Rights Council (HRC) and General Assembly (UNGA) in 2011, state that governments are obliged to take reasonable steps to ensure that companies and other non-state actors respect human rights, and that companies have a responsibility to respect human rights in their activities worldwide, including through due diligence and impact assessment.[36] Consideration of human rights impacts ought therefore to be a standard part of corporate practice.

However, the extent of corporate responsibilities is only patchily understood. This situation is changing, slowly and gradually, as businesses find it in their interests to take account of human rights impacts.[37] Increasingly, both national laws and investors' environmental, social and governance (ESG) or equivalent frameworks, plus civil society and public pressure, are obliging companies to give due regard to human rights. The European Commission's proposed directive

---

35 UN Office of the High Commissioner on Human Rights (undated), 'B-Tech Project', https://www.ohchr.org/en/business-and-human-rights/b-tech-project (accessed 12 Sep. 2022).
36 UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights,* https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.
37 Moynihan, H. and Alves Pinto, T. (2021), *The Role of the Private Sector in Protecting Civic Space*, Synthesis Paper, London: Royal Institute of International Affairs, https://www.chathamhouse.org/2021/02/role-private-sector-protecting-civic-space.

requiring mandatory human rights and environmental due diligence by larger companies based or active in the EU would be transformative and should herald a consistency of approach within the EU.[38]

Myth 5. 'Human rights are radical'

There are two dimensions to this particular myth: first, that – in line with popular news coverage – human rights are only relevant to extreme situations, such as the treatment of criminals, immigrants or terrorists. This view is plain wrong: human rights are about everyday protection from harm and discrimination for every adult and child, living free from state interference, and being provided with basic entitlements. In democracies, most people have a general, unspoken assumption that their human rights will be respected: for example, if arrested they will be treated with dignity; if prosecuted they will be granted a fair trial in a language they understand; or if voting their vote will be secret and will be counted fairly. Human rights routinely inform new legislation and policies, from data protection to social housing and social security. They are not often politically controversial. They are only newsworthy on the rare occasions when they are denied, or when they are portrayed as an obstacle to popular policies. The human rights law framework is not a radical philosophy, but a check and balance against discrimination or indignity in policy development.

The second dimension to this myth is a misconception that human rights are absolutist in nature: that, for example, they prohibit developments such as facial recognition technology. The desire for quick political soundbites in today's world encourages absolutist positions that can do human rights a disservice. The reality of human rights is more nuanced. For example, many civil society organizations currently assert that all facial recognition technology is contrary to human rights law.[39] But this is a shorthand for asserting that facial recognition as commonly configured (i.e. involving mass capturing and retention of personal data and potentially discriminatory judgements without regard to human rights considerations) is contrary to human rights law. In fact, human rights law does not lead to a conclusion that facial recognition, properly configured and constrained, should be banned where there are good reasons of safety or security for using it.[40] Rather, in this case as elsewhere, human rights law balances rights and interests to reach nuanced, subtle judgements.

---

38 European Commission (2022), *Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence*, COM(2022) 71 final (23.02.22), https://ec.europa.eu/info/sites/default/files/1_1_183885_prop_dir_susta_en.pdf.
39 For example, Liberty's website states that: 'Facial recognition technology…breaches everyone's human rights, discriminates against people of colour and is unlawful. It's time to ban it.' See Liberty (2022), 'Facial Recognition', https://www.libertyhumanrights.org.uk/fundamental/facial-recognition.
40 See, for example, Information Commissioner's Office (2021), *Information Commissioner's Opinion: The use of live facial recognition technology in public places*, https://ico.org.uk/media/for-organisations/documents/2619985/ico-opinion-the-use-of-lfr-in-public-places-20210618.pdf.

Myth 6. 'Human rights are vague'

There is a perception that human rights norms are too vague to guide AI. For example, some advocates of ethics argue that human rights are unable to provide guidance when values conflict.[41] These objections are largely unfounded. One strength of human rights law is its system for weighing competing rights and interests, whether the balance is to be struck between competing individual rights or with other collective or societal interests.[42]

One strength of human rights law is its system for weighing competing rights and interests.

Many human rights are framed in terms that make this balancing explicit. For example, Article 21 of the International Covenant on Civil and Political Rights states that the right of peaceful assembly shall be subject to no restrictions, 'other than those imposed in conformity with the law and… necessary in a democratic society in the interests of national security or public safety, public order (*ordre public*), the protection of public health or morals or the protection of the rights and freedoms of others'. In considering whether this right has been violated, the UN Human Rights Committee will consider first whether there has been an interference, then if so, whether that interference is lawful and both 'necessary for and proportionate to' one or more of the legitimate grounds for restriction listed in the article.[43] UN human rights bodies, national and regional courts have developed extensive jurisprudence on the appropriate balancing of rights and interests, balancing flexibility with predictability. These well-established, well-understood systems have repeatedly proven themselves capable of adaptation in the face of new policy tools and novel situations. For example, the European Court of Human Rights (ECtHR) recently developed new tests by which to assess bulk interception of online communications for intelligence purposes.[44]

The impact of AI is a novel but not insurmountable challenge, as emerging jurisprudence is already demonstrating. Indeed, one strength of international human rights law is its capacity to develop incrementally both as societal standards progress and in the face of new factual situations.[45]

Myth 7. 'Human rights get it wrong'

Some may consider that human rights protect the wrong values, apply protection in the wrong ways or are too rigid to apply to technological or social developments.

---

41 Canca, C. (2019), 'Why Ethics cannot be replaced by the Universal Declaration of Human Rights', UN University Our World, 15 August 2019, https://ourworld.unu.edu/en/why-ethics-cannot-be-replaced-by-the-universal-declaration-of-human-rights.
42 Yeung, Howes and Pogrebna (2019), 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight'.
43 UN Human Rights Committee (2020), *General Comment No. 37 on the right of peaceful assembly (Article 21),* para. 36.
44 *Big Brother Watch and others v UK* (ECtHR App no 58170/13).
45 See section 3.3 below.

For example, it has been suggested by some policymakers and academics that the individual right to privacy should be replaced or augmented by a concept of collective interest in appropriate handling of data that is sensitive to the interests of minority groups.[46] Group privacy may be a useful political concept in assessing appropriate limits of state or corporate power resulting from mass collection and processing of data.[47] But it cannot substitute for human rights law. Such claims underestimate the flexibility of human rights and its processes, including due diligence and human rights impact assessment, to secure the protection of human rights for all rather than just for those who claim infringement. The right to privacy is capable of evolution in light of competing interests, and enables a balance to be struck between privacy and the public interest in data-sharing and accessibility, while safeguarding the interests of groups categorized as such by AI by insistence on both freedom from discrimination and fairness and due process in decision-making. There may be scope for considering greater empowerment of data subjects[48] and/or group enforcement of rights; but it would be a rash move to abandon many years of judicial interpretation and scholarship, including concerns about the displacement of individual rights by group rights, by adding, or replacing them with, new legal constructs.

Myth 8. 'Human rights are organized around national models'

Human rights obligations are primarily owed by a state's government to people within that state's territory or jurisdiction. These jurisdictional limitations are under pressure: for example, UNGA has stressed that arbitrary surveillance and collection of personal data can violate various human rights, including when undertaken extraterritorially.[49] Regarding businesses, the corporate responsibility to respect human rights applies in respect of all individuals affected by a company's operations, regardless of location.[50] In practical terms, businesses should consider their human rights responsibilities towards everyone impacted by their work, in any country.

Myth 9. 'Human rights entail greater legal risk'

Human rights are legal rules, and so do entail accountability through courts and tribunals. But this accountability does not hinge on whether an organization pays attention to human rights, but on whether it is liable by reference to a rule of law.

---

46 For example, Mantelero (2016), 'Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection', *Computer Law & Security Review,* February 2016, 32(2), pp. 238–55, https://doi.org/10.1016/j.clsr.2016.01.014.
47 van der Sloot, B. (2017), 'Do groups have a right to protect their group interest in privacy and should they? Peeling the onion of rights and interests protected under Article 8 ECHR', in Taylor, L., Floridi, L. and van der Sloot, B. (2017), *Group Privacy: New Challenges of Data Technologies,* Cham: Springer, p. 223.
48 Wong, J., Henderson, T. and Ball, K. (2022), 'Data protection for the common good: Developing a framework for a data protection-focused data commons', *Data & Policy*, 4(e3), https://doi.org/10.1017/dap.2021.40.
49 UN General Assembly Resolution (2020), *The right to privacy in the digital age*, A/RES/75/176 (28 December 2020), preambular para. 24.
50 UN Office of the High Commissioner for Human Rights (2011), *Guiding Principles on Business and Human Rights*, principle 11.

Considering human rights will not place a company or government at greater risk from human rights claims. On the contrary, addressing human rights issues should help to protect against potential claims.

3.3 What human rights have to offer

Human rights law provides a means to define the harm that AI should avoid.[51] It places its focus on the interests of each individual and addresses the most pressing societal concerns of AI, including non-discrimination, fairness and privacy. It provides an excellent starting point by which to assess whether and to what extent AI is 'for good'. Economic and social rights offer a basis for considering societal distribution of AI's potential benefits.

Human rights offer a framework for regulating AI that is an existing system of international, regional and domestic law, commanding international legitimacy and a shared language across the world. This framework should be adopted in respect of AI, not only because of its intrinsic merit but because the current geopolitical stasis is likely to prevent effective multilateral cooperation on new normative frameworks. The focus of discussion should not be on whether human rights can or should be applied to AI, nor on potential alternatives, but on how the existing framework of human rights does apply in the field of AI. This is already the focus of international organizations at both regional and global level.[52]

Human rights crystallize a set of ethical values into international norms.[53] The system is not perfect, and was not created with AI in mind, but is a universally agreed blueprint for the protection of human values and the common good that has proven itself capable of adaptation to new circumstances. It avoids the need for fresh theoretical debates on the relative merits of different approaches. As a set of norms, human rights avoid the allegation – often levelled at ethics – of being vague and malleable enough to suit corporate interests.

Human rights are relatively clear. It is possible to list comprehensively the legally binding international, regional and domestic human rights obligations that apply in each country in the world. The meaning of those obligations is reasonably well-understood.[54]

The human rights approach has proved relatively successful over more than 70 years, developing incrementally with the benefit of several generations of academic input, governmental negotiation, civil society input and court

---

51 McGregor, L., Murray, D. and Ng, V. (2019), 'International Human Rights Law as a Framework for Algorithmic Accountability', *International & Comparative Law Quarterly*, 68(2), April 2019, https://doi.org/10.1017/S0020589319000046, pp. 324–27.
52 For example, the UN B-Tech Project, https://www.ohchr.org/en/business-and-human-rights/b-tech-project; and Council of Europe's Committee on Artificial Intelligence, https://www.coe.int/en/web/artificial-intelligence/cai.
53 'There is no conflict between ethical values and human rights, but the latter represent a specific crystallisation of these values that are circumscribed and contextualised by legal provision and judicial decisions'. Mantelero, A. and Esposito, S. (2021), 'An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems', *Computer & Security Law Review*, 2021, https://ssrn.com/abstract=3829759, p. 6.
54 UN Special Rapporteur on Extreme Poverty (2019), *Report on use of digital technologies in the welfare state*, A/74/493, https://digitallibrary.un.org/record/3834146?ln=en#record-files-collapse-header.

rulings from many parts of the world. It has evolved in tandem with societal development, its impact gradually increasing without meeting widespread calls for abandonment or radical change.

Human rights provide processes and accountability as well as principles

Human rights law is accompanied by a vast range of practical tools for implementation, political oversight and legal accountability that are absent from ethics.[55] Breaches of human rights entail legal as well as political avenues of redress. The international human rights framework includes a range of remedial mechanisms with practical effect, ranging from civil society advocacy through domestic and international courts, to scrutiny by UN bodies and other states. In many parts of the world, violations of rights by government may be challenged in court with legally binding effect – acting as an important constraint on state power.

As companies and governments already have human rights commitments, their use of AI will be scrutinized by human rights mechanisms in any case, including through claims made to domestic courts in the event of alleged breach. Human rights have already formed the basis for high-profile rulings on, for example, image databases[56] and uses of facial recognition technology.[57]

Human rights have international acceptance and legitimacy

International human rights law benefits from a higher degree of international acceptance and legitimacy than any other system of values. Governments in every continent know and understand the core human rights treaties. Every state is party to some of them, while some treaties have near-universal ratification. This remains the case, despite an apparently waning commitment to the universality of human rights in the rhetoric of certain countries.[58] Human rights have played a role, to a greater or lesser extent, in shaping the policies and activities of governments around the world.[59]

---

55 van Veen, C. and Cath, C., (2018), 'Artificial Intelligence: What's Human Rights Got to Do with It?', Data & Society: Points blog, 14 May 2018, https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5; Latonero, M. (2020), *AI Principle Proliferation as a Crisis of Legitimacy*, Carr Center Discussion Paper Series, Issue 2020-011, https://carrcenter.hks.harvard.edu/files/cchr/files/mark_latonero_ai_principles_6.pdf?m=1601910899.

56 For example, by the Canadian Office of the Privacy Commissioner (2021), 'Clearview AI's unlawful practices represented mass surveillance of Canadians, commissioners say', news release, 2 February 2021, https://www.priv.gc.ca/en/opc-news/news-and-announcements/2021/nr-c_210203/?=february-2-2021.

57 For example, the Marseille Administrative Tribunal ruled against the use of facial recognition technology at the entrances to French high schools in the *La Quadrature du Net* case, https://www.laquadrature.net/wp-content/uploads/sites/8/2020/02/1090394890_1901249.pdf.

58 For example, Position Paper of the People's Republic of China for the 77th Session of the United Nations General Assembly, 20 September 2022, http://geneva.china-mission.gov.cn/eng/dbdt/202209/t20220921_10768735.htm, section IV.

59 Latonero, M. (2020), *AI Principle Proliferation as a Crisis of Legitimacy*, Carr Center Discussion Paper Series, Issue 2020-011, https://carrcenter.hks.harvard.edu/files/cchr/files/mark_latonero_ai_principles_6.pdf?m=1601910899, p. 6.

UN processes affecting all states, such as the HRC's Universal Periodic Review and the UN treaty bodies' periodic examinations of states' compliance, entail that every UN member state engages with the international human rights architecture. Regional treaties that have strong local support reinforce these UN instruments in some parts of the world.[60] International human rights law has constitutional or quasi-constitutional status in many countries, notably in Europe, embedding it deep into systems of governance.[61] Civil society uses the human rights law framework as a basis for monitoring state and corporate activities worldwide.

International human rights law offers a degree of discretion to governments as to how they implement each right, within certain parameters.

This international legitimacy has given human rights a significant role in the production of internationally negotiated sets of AI governance principles. For example, the OECD AI Principles call on all actors to respect the rule of law, human rights and democratic values throughout the AI system life cycle.[62] As discussed previously, UNESCO's Recommendation on the Ethics of Artificial Intelligence names human rights and fundamental freedoms as the first of the 'values' around which it is crafted.[63] The Council of Europe's Committee on Artificial Intelligence (CAI) is working on a potential legal framework for the development, design and application of AI, based on the Council's standards on human rights, democracy and the rule of law.[64] Although the universality of human rights is increasingly contested, there is still, to a large degree, a global consensus on the continued relevance of long-agreed human rights commitments.

Human rights achieve a balance between universality and sensitivity to national contexts

International human rights law offers a degree of discretion to governments as to how they implement each right, within certain parameters. This flexibility is known as the 'margin of appreciation' in Europe, now enshrined in the preamble to the ECHR,[65] and has similar effect in the UN human rights system.[66] It varies according to the specific right in question and the impact of any interference: for example, human rights law offers governments no discretion in implementing bans on torture or slavery, but European human rights law permits governments a narrow

60 For example, ECHR; Inter-American Charter on Human Rights; African Charter on Human and People's Rights.
61 Yeung, Howes and Pogrebna (2019), 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight'.
62 Organisation for Economic Co-operation and Development (2019), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, Article 1.2(a).
63 UN Educational, Scientific and Cultural Organization (2021), *Recommendation on the Ethics of Artificial Intelligence*.
64 Council of Europe (2022), 'Inaugural Meeting of the Committee on Artificial Intelligence (CAI)'.
65 Protocol 15 to the ECHR, Article 1.
66 Shany, Y. (2018), 'All Roads Lead to Strasbourg?: Application of the Margin of Appreciation Doctrine by the European Court of Human Rights and the UN Human Rights Committee', *Journal of International Dispute Settlement*, 9(2), May 2018, pp. 180–98, https://doi.org/10.1093/jnlids/idx011.

margin of appreciation concerning general bans on protest, and a wider margin concerning whether governments choose to sanction protestors who intentionally disrupt ordinary life.[67]

Human rights are necessary but not sufficient for AI governance

International human rights law may not currently address all the potential harms to people caused by AI. But it is adaptable to new circumstances and changing social norms: the ECHR, for example, is 'a living instrument which… must be interpreted in light of present-day conditions.'[68] The UN secretary-general's High Level Panel on Digital Cooperation has called for an urgent examination of how human rights frameworks apply in the digital age.[69]

Human rights law may develop through new attention to existing rights. For example, the rights to freedom of thought and opinion are absolute. However, their parameters remain relatively unclear because they were largely taken for granted until challenged by the emergence of a technologically enabled industry of influence.[70] Further, new contexts may lead to new understandings and formulations of rights. For example, explainability and human involvement – commonly discussed elements of AI ethics – are not usually considered as elements of human rights, but might be found in existing requirements that individuals be provided with reasons for decisions made concerning them, and of the possibility of contesting those decisions and securing adequate remedies. The Council of Europe's work on a potential convention is likely to clarify the application of human rights to AI,[71] as human rights litigation is already beginning to do.[72]

The development of human rights law and its subsequent interpretation take time, yet technology moves quickly. Human rights in their current form, while essential, are not sufficient to act as an entire system for the ethical management of AI. Human rights should rather be the starting point for normative constraints on AI, the baseline to which new rights or further ethical guardrails might appropriately be added, including any ethical principles that businesses or other entities may choose to adopt.

The second half of this paper explores the contributions of human rights in detail and concludes by recommending practical actions to place human rights at the heart of AI governance.

---

67 European Court of Human Rights (2021), *Guide on Article 11 of the European Convention on Human Rights*, https://echr.coe.int/Documents/Guide_Art_11_ENG.pdf.
68 *Tyrer v United Kingdom*, ECtHR App No 5856/72, judgment of 25 April 1978, Series A No 26, para. 31.
69 UN Secretary-General's High-Level Panel on Digital Cooperation (2019), *The Age of Digital Interdependence* https://www.un.org/en/pdfs/DigitalCooperation-report-for%20web.pdf.
70 Jones (2019), *Online Disinformation and Political Discourse: Applying a Human Rights Framework*.
71 Council of Europe (2022), 'Inaugural Meeting of the Committee on Artificial Intelligence (CAI)'.
72 See Chapter 6.1 below.

04 Principles of AI governance: the contribution of human rights

Key principles of human rights law have an important role to play in determining AI governance standards.

There are three dimensions to AI governance: (i) the substantive standards, or principles, that the developers and implementers of AI should meet; (ii) the processes to ensure that substantive standards are met; and (iii) accountability and remedies for any breach of those standards.

In each of these dimensions, AI governance is immature because technology and its uses have developed much more rapidly than the rules constraining them. Human rights law offers baseline standards for all three dimensions.

4.1 Principles: the landscape

AI ethical principles from companies, civil society and intergovernmental organizations have proliferated in recent years,[73] causing more confusion than clarity through their overlapping nature, number and diversity.[74] There

73 The AI Ethics Guidelines Global Inventory lists over 165 sets of guidelines. AlgorithmWatch (2022), 'AI Ethics Guidelines Global Inventory', https://inventory.algorithmwatch.org.
74 Floridi, L. and Cowls, J. (2019), 'A Unified Framework of Five Principles for AI in Society', *Harvard Data Science Review*, 1.1, https://doi.org/10.1162/99608f92.8cd550d1.

are common themes such as data protection, understandability, transparency for accountability and tackling bias. But the precise meaning of each of these terms varies.[75] Some ethics principles identified – such as beneficence and non-maleficence[76] – are so abstract that they are not easily translatable for practical use in governance. There is no unifying theme between rival sets of ethical principles and there are debates on the representativeness of those principles, as most stem largely from Europe and North America, from separate corporate and national contexts, and from men.[77]

Some assert that, without unanimity as to what it entails, ethics offers a lexicon that can be used to give a veneer of respectability to any corporate activity. In the words of Philip Alston, 'as long as you are focused on ethics, it's mine against yours. I will define fairness, what is transparency, what is accountability. There are no universal standards.'[78]

4.2 Principles: Human rights law

To date, there are no international human rights treaties that specifically address the impact of AI,[79] but existing human rights laws apply to applications of AI. The former UN high commissioner for human rights, Michelle Bachelet, clarified that AI can have significant impacts on the implementation of many human rights, including privacy, health, education, freedom of movement, freedom of assembly and association, and freedom of expression.[80] Bachelet noted that inferences and predictions about individuals made by AI may profoundly affect not only those individuals' privacy but also their autonomy, and may raise issues regarding freedom of thought and opinion, freedom of expression, the right to a fair trial and other related rights.[81] Uses of faulty data may result in bias or discrimination,[82] as may faulty AI tools. Uses of AI in the criminal justice process may lead to violations of the rights to privacy, fair trial, freedom from arbitrary arrest and detention and even the right to life.[83]

While all rights are relevant, this section provides an overview of key rights that should form the basis of any safeguards for AI development.

---

75 Fjeld, J. et al. (2020), *Principled Artificial Intelligence*; Hagendorff (2020), 'The Ethics of AI Ethics'; Floridi and Cowls (2019), 'A Unified Framework of Five Principles for AI in Society'.
76 Floridi and Cowls (2019), 'A Unified Framework of Five Principles for AI in Society'.
77 Hagendorff (2020), 'The Ethics of AI Ethics'; Montreal AI Ethics Institute (2021), 'The Proliferation of AI Ethics Principles: What's Next?', https://montrealethics.ai/the-proliferation-of-ai-ethics-principles-whats-next.
78 UN Special Rapporteur on Extreme Poverty (2019), *Report on use of digital technologies in the welfare state*, A/74/493, https://digitallibrary.un.org/record/3834146?ln=en#record-files-collapse-header. See also Yeung, Howes and Pogrebna (2019), 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight', p. 3: 'Yet the vagueness and elasticity of the scope and content of "AI ethics" has meant that it currently operates as an empty vessel into which anyone (including the tech industry, and the so-called Digital Titans) can pour their preferred "ethical" content.'
79 Work is under way at the Council of Europe for a legal instrument on AI, by reference to the Council of Europe's standards on human rights, democracy and the rule of law. See Council of Europe (2022), 'Inaugural Meeting of the Committee on Artificial Intelligence (CAI)'.
80 United Nations High Commissioner for Human Rights (2021), *The Right to Privacy in the Digital Age*, A/HRC/48/31, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/249/21/PDF/G2124921.pdf.
81 Ibid., para. 17.
82 Ibid., para. 19.
83 Ibid., para. 24.

4.2.1 Privacy

The challenges presented by AI

AI is having a huge impact on privacy and data protection. Far more information about individuals is collated now than ever before, increasing the potential for exploitation. A new equilibrium is needed between the value of personal data for AI on the one hand and personal privacy on the other. There are two parallel challenges to overcome: (i) AI is causing, and contributing to, significant breaches of privacy and data protection; and (ii) use of extensive personal data in AI decision-making and influencing is contributing to an accretion of state and corporate power.

Examples of breaches of privacy and data protection include:

- AI's requirement for data sets may create an incentive for companies and public institutions to share personal data in breach of privacy requirements. For example, in 2017, a UK health trust was found to have shared the data of 1.6 million patients with Google's DeepMind, without adequate consent from the patients concerned.[84]
- AI may facilitate the harvesting of personal data without adequate consent. Between 2013 and 2018, Cambridge Analytica collated personal data of up to 87 million Facebook users without their knowledge or consent for use in political advertising.[85]
- The practice of using publicly available images to create AI facial recognition databases raises major privacy concerns. Projects such as Exposing.ai aim to highlight the privacy implications of extant large facial recognition datasets.[86] Some large companies, including Microsoft and Facebook, have closed their facial recognition operations.[87] Clearview AI's provision of facial recognition technology for law enforcement purposes – via a database of 10 billion images gleaned from the internet – has been found in breach of privacy laws in several countries, including Australia, Canada, France and the UK.[88]
- AI lends itself to bulk interception and assessment of online communications. In 2021, the ECtHR found that the UK's former regime for bulk interception, using digital and automated methods, lacked necessary end-to-end safeguards for compliance with privacy rights.[89]

---

84 BBC News (2017), 'Google DeepMind NHS app test broke UK privacy law', 3 July 2017, https://www.bbc.co.uk/news/technology-40483202.
85 Information Commissioner's Office (2018), *Investigation into the use of data analytics in political campaigns*, https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf.
86 Harvey, A. and LaPlace, J. (2021), 'Exposing.ai', 1 January 2021, https://exposing.ai (accessed 12 Sep. 2022).
87 Microsoft withdrew its MS Celeb database in 2019: Computing News (2019), 'Microsoft withdraws facial recognition database of 100,000 people', 6 June 2019, https://www.computing.co.uk/news/3076968/microsoft-withdraws-facial-recognition-database-of-100-000-people. Meta announced in November 2021 that it was shutting down Facebook's facial recognition system: Meta (2021), 'An update on our use of face recognition', 2 November 2021, https://about.fb.com/news/2021/11/update-on-use-of-face-recognition.
88 Lomas, N. (2021), 'France latest to slap Clearview AI with order to delete data', TechCrunch, 16 December 2021, https://techcrunch.com/2021/12/16/clearview-gdpr-breaches-france.
89 *Big Brother Watch and others v UK* (ECtHR App no 58170/13).

— 'Smart' devices, such as fridges and vehicles, may not only collate data on users to improve performance, but also to sell to third parties. If not properly secured, such devices may also expose users to surveillance by hackers. In 2017, for example, the German authorities withdrew the 'My Friend Cayla' doll from sale over fears that children's conversations could be listened to via Bluetooth.[90]

AI impacts privacy in several ways. First, its thirst for data creates compelling reasons for increased collection and sharing of data, including personal data, with the aim of improving the technology's operation. Second, AI may be used to collate data, including that of a sensitive, personal nature, for purposes of surveillance. Third, AI may be used to develop profiles of individuals that are then the basis of decisions on matters fundamental to their lives – from healthcare to social benefits, to employment to insurance provision. As part of this profiling, AI may infer further, potentially sensitive information about individuals without their knowledge or consent, such as conclusions on their sexual orientation, relationship status or health conditions. Finally, AI may make use of personal data to micro-target advertising and political messaging, to manipulate and exploit individual vulnerabilities, or even to facilitate crimes such as identity theft.

International human rights law

The human right to privacy currently entails that any processing of personal data should be fair, lawful and transparent, based on free consent or another legitimate basis laid down in law. Data should only be held for a limited period and for specific purposes, with those purposes not to be lightly changed. Data should be held securely, and sensitive personal data should enjoy heightened protection. Privacy entails that individuals should know that their personal data has been retained and processed, and that they have a right both to rectify or erase their personal data and to limit how it is used. Privacy further entails that individuals must not be exposed to mass surveillance or unlimited profiling. Personal data should not be transferred, particularly overseas, unless similar standards will be upheld by the recipient of that data.[91]

Human rights law is already the widely accepted basis for most legislation protecting privacy. The EU's General Data Protection Regulation (GDPR) is founded on the right to protection of personal data in Article 8(1) of the EU Charter of Fundamental Rights – this is an aspect of the right to privacy in earlier human rights treaties. Privacy and data protection is one of the European Commission's Seven Principles for Trustworthy AI, while most statements of AI principles include a commitment to privacy.[92]

Application of human rights law to the challenges of AI

With the development of AI, it is becoming apparent that changes need to be made to the contours of the right to privacy.

---

90 BBC News (2017), 'German parents told to destroy Cayla toys over hacking fears', 17 February 2017, https://www.bbc.co.uk/news/world-europe-39002142.
91 United Nations High Commissioner for Human Rights (2018), *The Right to Privacy in the Digital Age*, A/HRC/39/29, https://www.ohchr.org/en/documents/reports/ahrc3929-right-privacy-digital-age-report-united-nations-high-commissioner-human.
92 35 of the 36 statements of AI principles reviewed by Fjeld et al. included this commitment: Fjeld et al. (2020), *Principled Artificial Intelligence*.

There is growing awareness of the tension between privacy's requirement to restrict flows of personal data on the one hand, and economic and commercial arguments in favour of free flow on the other. There are many sound reasons for improved data accessibility: fostering developments in AI innovation; facilitating increased use of AI; and preventing data restrictions from distorting markets or acting as a barrier to competition and innovation.

Privacy should not be viewed as static: it is flexible enough to adapt and develop [...] in light of rapidly changing technological and social conditions.

Privacy should not be viewed as static: it is flexible enough to adapt and develop, through new legislation or through judicial interpretation, in light of rapidly changing technological and social conditions. Individual privacy remains vital to ensuring that individuals do not live in a surveillance state, and that individuals retain control over their own data and by whom and how it is seen and used. This is critical at a time when the value of privacy is being steadily and unconsciously diluted.

The human right to privacy should be used to resolve competing interests in an AI-dominated world – whether those interests are commercial, individual or technical. For example, rather than privacy impeding the transfer of anonymized data for use in AI data sets, the balancing between rights and interests allowed by the human right to privacy could be used to set appropriate limits on data-profiling and micro-targeting.

4.2.2 Equality: discrimination and bias

The challenges presented by AI

Because AI generally operates by applying rules to the treatment of people, rather than by assessing each individual on their merits, it carries significant risks of embedding discrimination, as the rules that it applies may distinguish between people, directly or indirectly, by reference to protected characteristics. Indeed, examples of such bias and discrimination in the use of AI abound:

- In 2015, researchers found that female job seekers were much less likely than males to be shown adverts for highly paid jobs on Google.[93]
- In 2016, researchers found that an algorithm used to determine offenders' risk of recidivism often overstated the risk that black defendants would re-offend, and understated the risk of reoffending by white defendants.[94]

---

93 Gibbs, S. (2015), 'Women less likely to be shown ads for high-paid jobs on Google, study shows', *Guardian*, 8 July 2015, https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study.
94 Larson, J. et al. (2016), 'How We Analyzed the COMPAS Recidivism Algorithm', *ProPublica*, 23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm; see also *State of Wisconsin v Eric L Loomis* (2016) WI 68, 881 N.W.2d 749.

- In 2017, Amazon abandoned its automated recruitment platform, built on observing patterns in applicant CVs over the previous years, having been unable to prevent it from discriminating on the basis of gender or from making other inappropriate recommendations.[95]
- In 2018, Immigration New Zealand suspended its use of data-profiling, which had been predicting likely healthcare costs and criminality of immigrants on the basis of demographics including age, gender and ethnicity.[96]
- In 2019, researchers found that AI widely used to allocate healthcare in US hospitals was systematically discriminating against black people, by referring them on to specialized care programmes less frequently than white people. The algorithm was predicting future healthcare costs as a proxy for illness, using past costs for individuals in similar situations. This failed to take account of the fact that less money had been spent historically on caring for black patients.[97]
- In 2020, the Austrian public employment service (AMS) began using an algorithm that enabled it to classify jobseekers according to their likelihood of successful re-employment. The algorithm has been criticized for discriminating on the basis of gender, disability and other factors, and for intersectional discrimination.[98] AMS has suspended use of the algorithm pending the outcome of legal challenges.[99]

AI makes it difficult to assess whether discrimination has occurred. An individual usually becomes aware of discrimination by comparing their treatment, or its outcome, with that of other people. But when complex AI is used to make each individual a personalized offer (for example, on social security payments) or decision (for example, on school or college entry), that individual may have no means of knowing what criteria were used, nor how their result differs from others. Consequently, individuals may not know, or have any accessible way of finding out, whether they have been disadvantaged or how.[100]

AI developers have learned from past problems and gone to considerable lengths to devise systems that promote equality as much, or more, than human decision-making.[101] Nonetheless, several features of AI systems may cause them to make biased decisions. First, AI systems rely on training data to train the decision-making algorithm. Any imbalance or bias in that training data is likely then to be replicated and become exaggerated in the AI system. If the training data is taken from the real world, rather than artificially generated, AI is likely to replicate and exaggerate any

---

95 Dastin, J. (2018), 'Amazon scraps secret AI recruiting tool that showed bias against women', Reuters, 11 October 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
96 Bonnett, G. (2018), 'Immigration NZ using data system to predict likely troublemakers', *RNZ News*, 5 April 2018, https://www.rnz.co.nz/news/national/354135/immigration-nz-using-data-system-to-predict-likely-troublemakers.
97 Obermeyer, Z. et al. (2019), 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, 25 October 2019, 366(6464), pp. 447–553, https://doi.org/10.1126/science.aax2342.
98 Allhutter, D. et al. (2020), 'Algorithmic profiling of job seekers in Austria: how austerity politics are made effective', *Frontiers in Big Data*, 21 February 2020, https://doi.org/10.3389/fdata.2020.00005.
99 Der Standard (2022), '"Zum In-die-Tonne-Treten": Neue Kritik am AMS-Algorithmus' ["To Be Thrown In The Bin": New Criticism of the AMS Algorithm'], 28 April 2022, https://www.derstandard.at/story/2000135277980/neuerliche-kritik-am-ams-algorithmus-zum-in-die-tonne-treten.
100 Obermeyer et al. (2019), 'Dissecting racial bias in an algorithm used to manage the health of populations'.
101 For example, HireVue, an AI recruitment tool used by some large companies, claims to '[i]ncrease diversity and mitigate bias' by finding a wider candidate pool, evaluating objectively and consistently, and helping to avoid unconscious bias. See HireVue (2022), https://www.hirevue.com/employment-diversity-bias.

bias already present in society. Second, AI systems rely on the instructions given to them, as well as their own self-learning. Any discrimination or bias deployed by the designer risks being replicated and exaggerated in the AI system. Third, AI systems operate within a context: an AI system will lead to bias if it is deployed within the context of social conditions that undermine enjoyment of rights by certain groups.[102] Without human involvement, AI is currently unable to replicate contextual notions of fairness.

International human rights law

Human rights law provides standards of equality and non-discrimination by which to assess AI. It requires that all individuals' rights be respected and ensured 'without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status'.[103] The law entails prohibitions against not just direct discrimination (i.e. treating people differently on prohibited grounds), but indirect discrimination (i.e. treating people the same, but in a way that puts people from a protected group at a disadvantage without an objective justification) and structural discrimination (i.e. creating structural conditions in society that prevent all groups from accessing the same opportunities). Acknowledging that equality does not always mean treating everyone the same, discrimination law provides structured tests for assessing and preventing unlawful treatment.

This ban on discrimination has formed the basis for well-developed understandings of, and jurisprudence on, non-discrimination in both the public and private sectors. Human rights law obliges governments both to ensure there is no discrimination in public sector decision-making and to protect individuals against discrimination in the private sector. Human rights law does not forbid differential treatment that stems from factors other than protected characteristics, but such treatment must meet standards of fairness and due process in decision-making (see below).

Application of human rights law to the challenges presented by AI

Human rights practitioners are accustomed to considering the prohibition of discrimination by reference to well-established tests, and to resolving tensions between non-discrimination and other rights like freedom of speech. Adopting the standards that are well established and internationally accepted in human rights law minimizes the need for fresh debates on highly contested concepts in ethics (what is 'justice'? what is 'fairness'?).[104] Further, it avoids the risk of confusion from the imposition of parallel, non-human rights standards of discrimination specifically in the field of AI.

---

102 Wachter, S., Mittelstadt, B. and Russell, C. (2020), 'Why Fairness Cannot be Automated: Bridging the Gap between EU Non-Discrimination Law and AI', *Computer Law & Security Review*, 41(2021): 105567, https://ssrn.com/abstract=3547922.
103 International Covenant on Civil and Political Rights, Article 2(1). Some non-discrimination laws forbid discrimination in all circumstances, rather than merely in the implementation of rights: see Protocol 12 to the European Convention on Human Rights and Articles 20 and 21 of the European Charter of Fundamental Rights.
104 For example on justice, Floridi, L. et al. (2018), 'AI4People – An Ethical Framework for a Good AI Society', *Minds and Machines*, 28, pp. 689–707, https://doi.org/10.1007/s11023-018-9482-5; Bartneck, C. et al. (2021), *An Introduction to Ethics in Robotics and AI*, Springer Briefs in Ethics, p. 33.

International human rights law does not simply require governments to ban discrimination in AI. As the UN special rapporteur on contemporary forms of racism has observed, human rights law also requires governments to deploy a structural understanding of discrimination risks from AI. To combat the potential for bias, the tech sector would benefit from more diversity among AI developers, more guidance on bias detection and mitigation and the collection and use of data to monitor for bias, and more leadership by example from the public sector.[105] AI developers and implementers must consider holistically the impact of all algorithms on individuals and groups, rather than merely the impact of each algorithm on each right separately.[106] Algorithms should be reviewed regularly to ensure that their results are not discriminatory, even though obtaining data for comparison purposes may be challenging.[107] Vigilance is needed to ensure that other factors are not used as proxies for protected characteristics – for example, that postcode is not used as a proxy for ethnic origin.

Adopting well-established and internationally accepted standards in human rights law minimizes the need for fresh debates on highly contested concepts in ethics.

Legislators, regulators (such as the UK's Equality and Human Rights Commission) and courts need to consider the methodology for ensuring and overseeing compliance with the right to non-discrimination with regard to AI. New tools may be necessary to detect discrimination, as AI systems operate differently and are generally more opaque than non-AI decision-making processes. To be able to review the operation of AI effectively, the law and the courts may have to take more account of statistical method as well as context, while also adopting more standardized thresholds where possible and appropriate.[108] In parallel, AI developers need to ensure that automated decision-making matches its human equivalent by developing capacity to take account of a rich complexity of factors relevant to the circumstances of the individual. Legal and technical communities should work together to find adequate ways of reducing discrimination in algorithmic systems, including by embedding transparency and contextual approaches.

105 Centre for Data Ethics and Innovation (2020), *Review into Bias in Algorithmic Decision-Making*, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf, pp. 9–10.
106 McGregor, L., Murray, D. and Ng, V. (2019), 'International Human Rights Law as a Framework for Algorithmic Accountability', *International & Comparative Law Quarterly*, 68(2), April 2019, https://doi.org/10.1017/S0020589319000046, p. 326.
107 Centre for Data Ethics and Innovation (2020), *Review into Bias in Algorithmic Decision-Making*, pp. 9–10.
108 Wachter, S., Mittelstadt, B. and Russell, C. (2020), 'Why Fairness Cannot be Automated: Bridging the Gap between EU Non-Discrimination Law and AI', *Computer Law & Security Review*, 41 (2021): 105567, https://ssrn.com/abstract=3547922.

4.2.3 Autonomy

The challenges presented by AI

AI poses two principal risks to autonomy. First, empathic AI[109] is developing the capacity to recognize and measure human emotion as expressed through behaviour, expressions, body language, voice and so on.[110] Second, it is increasingly able to react to and simulate human emotion, with the aim of generating empathy from its human users. Empathic AI is beginning to appear in a multitude of devices and settings, from games and mobile phones, to cars, homes and toys, and across industries including education, insurance and retail. Research is ongoing as to how AI can monitor the mental[111] and physical health of employees.[112]

Some empathic AI has clear benefits. From 2022, EU law requires that new vehicles incorporate telemetrics for the detection of drowsiness and distraction in drivers.[113] Besides the obvious safety benefits for drivers and operators of machinery, empathic AI offers assistive potential (particularly for disabled people) and prospects for improving mental health. Other possible enhancements to daily lives range from recommendations for cures to ailments to curated music-streaming.[114]

However, empathic AI also carries major risks. The science of emotion detection and recognition is still in development, meaning that, at present, any chosen labelling or scoring of emotion is neither definitive nor necessarily accurate. Aside from these concerns, empathic AI also raises significant risks of both surveillance and manipulation. The use of emotion recognition technology for surveillance is likely to breach the right to privacy and other rights – for example, when used to monitor employee or student engagement or to identify criminal suspects.[115] More broadly, monitoring of emotion, as of all behaviour, is likely to influence how people behave – potentially having a chilling effect on the freedoms of expression, association and assembly, and even of thought.[116] This is particularly the case where access to rights and benefits is made contingent on an individual meeting standards of behaviour, as for instance in China's 'social credit' system.[117]

Regarding manipulation, empathic AI blurs the line between recommendation and direction. Algorithms may influence individuals' emotions and thoughts,

---

109 Also known as 'emotion AI', 'emotional AI', and 'affective computing' (a term coined by Rosalind Picard in her 1995 book on the topic). One example is sentiment analysis, which entails the assessment of text (such as customer feedback and comments) and, increasingly, of images (of people, objects or scenes) for emotional tone.
110 For an overview and research in this field, see Emotional AI Lab (undated), www.emotionalai.org.
111 For example, Lewis, R. et al. (2022), 'Can a Recommender System Support Treatment Personalisation in Digital Mental Health Therapy?', MIT Media Lab, 21 April 2022, https://www.media.mit.edu/publications/recommender-system-treatment-personalisation-in-digital-mental-health.
112 Whelan, E. et al. (2018), 'How Emotion-Sensing Technology Can Reshape the Workplace', MIT Sloan Management Review, 5 February 2018, https://sloanreview.mit.edu/article/how-emotion-sensing-technology-can-reshape-the-workplace.
113 General Safety Regulation, Regulation (EU) 2019/2144 of the European Parliament and of the Council, https://eur-lex.europa.eu/eli/reg/2019/2144/oj.
114 For a discussion of the potential of empathic AI, see McStay, A. (2018), *Emotional AI: The Rise of Empathic Media*, London: SAGE Publications Ltd, chap. 1.
115 Article 19 (2021), *Emotional Entanglement: China's emotion recognition market and its implications for human rights,* January 2021, https://www.article19.org/wp-content/uploads/2021/01/ER-Tech-China-Report.pdf.
116 UN Special Rapporteur on Freedom of Religion or Belief (2021), *Freedom of Thought*, A/76/380 (October 2021), https://undocs.org/Home/Mobile?FinalSymbol=A%2F76%2F380&Language=E&DeviceType=Desktop&LangRequested=False, para. 54.
117 See the discussion of China's social credit system in Taylor, E., Jones, K. and Caeiro, C. (2022), 'Technical Standards and Human Rights: The Case of New IP', in Sabatini, C. (2022), *Reclaiming human rights in a changing world order*, Washington, DC and London: Brookings Institution Press and Royal Institute of International Affairs, pp. 185–215.

and the decisions they make, without them being aware.[118] The distinction between acceptable influence and unacceptable manipulation has long been blurred. At one end of the spectrum, nudge tactics such as tailored advertising and promotional subscriptions are commonly accepted as marketing tools. At the other, misrepresentation and the use of fake reviews are considered unacceptable and attract legal consequences. Between those extremes, the boundaries are unclear.

Retail and other commercial sectors are increasingly harnessing empathic AI technology. For example, just as advertising has long sought to take advantage of mood and feeling to promote sales, micro-targeting could be taken a step further by including emotion detection as one of its parameters, with the aim of persuading an individual to book a holiday or sign up for a therapy class, among other things. There are currently no parameters by which to assess the acceptable limits of influence, even as persuasive tactics edge further towards manipulation.

In social media, too, AI offers potential for emotional manipulation, not least when it comes to politics. In particular, the harnessing of empathic AI exacerbates the threat posed by campaigns of political disinformation and manipulation. AI use to harness emotion for political ends has already been widely reported. This includes the deployment of fake or distorted material, often micro-targeted, to simulate empathy and inflame emotions.[119] Regulation and other policies are now being targeted at extreme forms of online influence,[120] but the parameters of acceptable behaviour by political actors remain unclear.

Empathic AI could have major impacts on all aspects of life. Imagine, for example, technology that alters children's emotional development, or that tailors career advice to young people in an emotionally empathic manner that appears to expand but actually has the effect of limiting choice. Vulnerable groups, including minors and adults with disabilities, are particularly at risk. Researchers of very large language models have argued for greater consideration of the risks of human mimicry and abuse of empathy they create.[121]

The draft EU Artificial Intelligence Act would ban the clearest potential for manipulation inherent in AI by prohibiting AI that deploys subliminal techniques to distort people's behaviour in a manner that may cause them 'physical or psychological harm'.[122] The Act would also limit the uses of individual 'trustworthiness' profiling. As most empathic AI involves the use of biometric

---

118 Council of Europe (2019), Declaration by the Committee of Ministers on the Manipulative Capabilities of Algorithmic Processes, Decl(13/02/2019)1.
119 Jones (2019), *Online Disinformation and Political Discourse: Applying a Human Rights Framework*.
120 For example, European Democracy Action Plan and related legislation: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (2020), *On the European Democracy Action Plan*, COM/2020/790 final https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_en#what-is-the-european-democracy-action-plan. In the UK, the National Security Bill, clauses 13 and 14 would criminalize foreign interference, while the government has announced its intention to make foreign interference a prioritized offence for the purposes of the Online Safety Bill.
121 Bender, E. et al. (2021), 'On the Dangers of Stochastic Parrots: Can Language Models be Too Big?', event, *FAccT 2021*, 3–10 March 2021, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922; Bender, E. (2022), 'Human-like Programs Abuse Our Empathy – even Google Engineers Aren't Immune', *Guardian*, 14 June 2022, https://www.theguardian.com/commentisfree/2022/jun/14/human-like-programs-abuse-our-empathy-even-google-engineers-arent-immune.
122 Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, Article 5.

data, it is likely to be subject to the Act's enhanced scrutiny for 'high-risk' AI. However, empathic AI that operates on an anonymous basis may not be covered.

International human rights law

As well as privacy, human rights law protects autonomy. It protects the right to freedom of thought and the right to hold opinions without interference, as well as the better-known and -understood rights to freedom of expression, freedom of assembly and association, and freedom of conscience and religion. The EU Charter of Fundamental Rights also protects the right to 'mental integrity'. Prior to recent technological developments, the rights to freedom of thought and opinion were underexplored. Further guidance is now emerging: for example, the UN special rapporteur on freedom of religion or belief has recently issued guidance on freedom of thought.[123]

Children's rights merit special consideration in this area. In addition to questions over privacy and the ability of minors to give consent when providing personal data, the UN Committee on the Rights of the Child has called for practices that rely on neuromarketing and emotional analytics to be prohibited from direct or indirect engagement with children,[124] and for states to prohibit manipulation or interference with the child's right to freedom of thought and belief through emotional analytics and interference.[125]

Application of human rights law to the challenges presented by AI

There are considerable concerns about the extent to which emotion recognition, capture and simulation may infringe human rights, in ways that are not necessary or proportionate to perceived benefits.

At present, challenges to autonomy are generally viewed through the prism of privacy and data protection. While this enables consideration of the impacts of surveillance, it is not a sufficient framework by which to consider issues of manipulation. Empathic AI can still be effective without capturing personal data – examples include billboards that adapt their advertising according to the reactions of people walking past, stores that adapt their advertising and marketing after capturing shoppers' reactions in real time or bots that reflect unnamed users' emotions in order to influence their decision-making.

Initiatives to set limits on simulated empathy, such as the technical standard under development by the IEEE,[126] ought to take account of the absolute nature of the rights to freedom of opinion and freedom of thought, as well as the right to mental integrity and the rights of the child. Further legislative and judicial consideration is needed to establish precisely what constraints human rights law imposes on potentially manipulative uses of AI, and precisely what safeguards it imposes to prevent the erosion of autonomy.

---

123 UN Special Rapporteur on Freedom of Religion or Belief (2021), *Freedom of Thought*, A/76/380 (October 2021), https://undocs.org/Home/Mobile?FinalSymbol=A%2F76%2F380&Language=E&DeviceType=Desktop&LangRequested=False, paras 68–72.
124 UN Committee on the Rights of the Child (2021), General Comment No. 25 on children's rights in relation to the digital environment, CRC/C/GC/25, 2 March 2021, para. 42.
125 Ibid., para. 62.
126 IEEE 7000-P7014, Empathic Technology Working Group on a Standard for ethical considerations in emulated empathy in autonomous and intelligent systems.

Meanwhile, some are reaching their own conclusions on empathic AI. For example, a coalition of prominent civil society organizations has argued that the EU's Artificial Intelligence Act should prohibit all emotion recognition AI, subject to limited exceptions for health, research and assistive technologies.[127] In June 2022, Microsoft announced that it would phase out emotion recognition from its Azure Face API facial recognition services. In that announcement, Microsoft noted the lack of scientific consensus on the definition of 'emotions', challenges of generalizations across diverse populations, and privacy concerns as well as awareness of potential misuse of the technology for stereotyping, discrimination or unfair denial of services.[128]

4.2.4 Equality: implementation of economic and social rights

International human rights law protects a wide range of economic and social rights, and provides an anchor for sustainable development.[129] Just as AI offers opportunities to achieve implementation of the SDGs, so it offers significant potential to improve the implementation of rights such as those to education, health, social security and work. Equality is key to achieving this potential: not just through the avoidance of discrimination, but through AI that benefits all communities and through the provision of equal opportunity for all in accessing the benefits. Failure to realize such opportunities risks not only entrenching but exacerbating current social divisions.

Ideally, such provision would begin with research into AI technologies that would help to implement the SDGs, and funding for the development and rollout of those technologies. The challenges are to incentivize developments that benefit all communities, as well as those that are most profitable; and to ensure that no AI systems operate to the detriment of vulnerable communities.

4.2.5 Fairness and due process in decision-making

AI decision-making brings a risk that the 'computer says no' in respect of significant life decisions, without possibility of review or challenge. Aside from discrimination, this also raises questions as to fairness of process and quality of decision-making in AI systems. It concerns both whether the use made of AI to reach the decision was fair, and whether AI reached or contributed to a fair decision in the specific case – and if not, what the recourse might be.

In making decisions, AI may segment people by reference to a wide range of factors and without consideration as to whether segmentation is appropriate in the particular case. These factors may be unrelated to the decision in question,

---

127 Joint Civil Society Amendments to the Artificial Intelligence Act (2022), *Prohibit Emotion Recognition in the Artificial Intelligence Act*, May 2022, https://www.accessnow.org/cms/assets/uploads/2022/05/Prohibit-emotion-recognition-in-the-Artificial-Intelligence-Act.pdf.
128 Bird, S. (2022), 'Responsible AI investments and safeguards for facial recognition', Microsoft *Azure* blog, 21 June 2022, https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition.
129 UN Office of the High Commissioner on Human Rights (undated), 'OHCHR and the 2030 Agenda for Sustainable Development', https://www.ohchr.org/en/sdgs.

but decisions that treat some people unfairly in comparison to others may still result. For example, if a travel insurance provider were to double the premiums offered to people who had opted out of receiving unsolicited marketing material, it would not be discriminating on the basis of a protected characteristic. Its decision-making process would however be biased against those who have opted out.

Where an individual's human rights are affected by a decision made by a public authority, they should be able to seek remedy[130] and will usually be able to challenge the decision in public law – for example, by way of judicial review. Decision-making processes need to be sufficiently transparent to enable such review. Individuals should know who the decision-maker is, the factors on which the decision is made and be able to verify the accuracy of any personal data used in the process. There should be adequate human involvement or oversight – while acknowledging that human involvement may not be essential in every case and is not necessarily a failsafe.[131]

International human rights law stipulates requirements for fairness in legal proceedings. Public and private law bases of challenge to decisions commonly reflect these requirements, and they can provide the basis for guidelines on minimum standards for transparency, human control and accountability through possibility of review for all AI activities.

4.2.6 Other rights

AI, used in different contexts, may have serious implications for the full range of human rights.

For example, the use of AI for content curation and moderation in social media may affect the rights to freedom of expression and access to information. The use of analytics to contribute to decisions on child safeguarding, meanwhile, may affect the right to family life.[132] The use of facial recognition technology risks serious impact on the rights to freedom of assembly and association, and even on the right to vote freely. In extreme cases – for example, in weapons for military use – AI risks undermining the right to life and the right to integrity of the person if not closely circumscribed. In each of these areas, existing human rights can form the basis for safeguards delimiting the appropriate scope of AI activity.

---

130 International Covenant on Civil and Political Rights, Art. 2(3); European Convention on Human Rights, Art. 13.
131 In the data protection context, there is pressure to change Article 22 of GDPR, which currently requires that decisions with legal or similarly significant effects for individuals, using their personal data, shall not be based solely on automated processing.
132 Anning, S. (2022), 'The Interplay of Explicit and Tacit Knowledge With Automated Systems for Safeguarding Children', *techUK Industry Views* blog, 21 March 2022, https://www.techuk.org/resource/the-interplay-of-explicit-and-tacit-knowledge-with-automated-systems-for-safeguarding-children.html.

05 Processes of AI governance: the contribution of human rights

Regulators and companies should follow human rights process requirements as they devise and implement AI governance processes.

5.1 Processes: the landscape

The processes that governments and companies should follow in order to meet AI governance standards are evolving rapidly.

5.1.1 Regulation

Governments are increasingly considering cross-sectoral regulation of AI on the basis that statutory obligations would help create a level playing field for safe and ethical AI and bolster consumer trust, while mitigating the risk that pre-AI regulation applies to AI in haphazard fashion.[133] The EU is furthest along in this process, with its draft Artificial Intelligence Act that would ban the highest-risk

133 In the UK, regulators have established the Digital Regulation Cooperation Forum to facilitate a joined-up approach to technology regulation. In the US, the Federal Trade Commission has explained how it stands ready to enforce existing legislation – including the Federal Trade Commission Act, the Fair Credit Reporting Act, and the Equal Credit Opportunity Act – against bias or other unfair outcomes in automated decision-making. See Jillson, E. (2021), 'Aiming for truth, fairness, and equity in your company's use of AI', Federal Trade Commission *Business Blog*, 19 April 2021, https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai.

forms of AI and subject other 'high risk' AI to conformity assessments. In the US, Congress is considering a draft Algorithmic Accountability Act.[134] The British government, having considered the case for cross-cutting AI regulation, has recently announced plans for a non-statutory, context-specific approach that aims to be pro-innovation and to focus primarily on high-risk concerns.[135]

While the British government, among others, has expressed concern that general regulation of AI may stifle innovation, many researchers and specialists make the opposite argument.[136] Sector-specific regulation may not tackle AI risks that straddle sectors, such as the impact of AI in workplaces. Well-crafted regulation should only constrain undesirable activity, and should provide scope for experimentation without liability within its parameters, including for small companies. Moreover, it is argued that responsible businesspeople would rather operate in a marketplace regulated by high standards of conduct, with clear rules, a level playing field and consequent consumer trust, than in an unregulated environment in which they have to decide for themselves the limits of ethical behaviour. Most decision-makers in industry want to do things the right way and need the tools by which to do so.

Without… clear standards and external involvement or accountability, there is a risk of 'ethics-washing' rather than genuine mitigation of risks.

In addition to regulating AI itself, there are also calls for regulation to ensure that related products are appropriately harnessed for the public good. For example, the UK-based Ada Lovelace Institute has called for new legislation to govern biometric technologies.[137] Similarly, there is discussion of regulation of 'digital twins' – i.e. computer-generated digital facsimiles of physical objects or systems – to ensure that the vast amounts of valuable data they generate is used for public good rather than for commercial exploitation or even public control.[138]

Some sector-specific laws are already being updated in light of AI's expansion. For example, the European Commission's proposal to replace the current Consumer Credit Directive aims to prohibit discrimination and ensure accuracy, transparency and use of appropriate data in creditworthiness assessments, with a right to human review of automated decisions.[139] An analysis of legislation

---

134 H.R. 6580 – Algorithmic Accountability Act of 2022, https://www.congress.gov/bill/117th-congress/house-bill/6580/text.
135 UK Government (2022), *Establishing a pro-innovation approach to regulating AI*, Policy Paper, 20 July 2022, https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement.
136 See, for example, Ada Lovelace Institute (2021), 'Regulate to innovate', 29 November 2021, https://www.adalovelaceinstitute.org/report/regulate-innovate.
137 Chang, M. (2022), 'Countermeasures: the need for new legislation to govern biometric technologies in the UK', London: Ada Lovelace Institute, 29 June 2022, https://www.adalovelaceinstitute.org/report/countermeasures-biometric-technologies.
138 See, for example, Centre for Digital Built Britain (2018), *The Gemini Principles*, Cambridge: University of Cambridge, https://www.cdbb.cam.ac.uk/system/files/documents/TheGeminiPrinciples.pdf.
139 European Commission (2021), *Proposal for a Directive of the European Parliament and of the Council on Consumer Credits*, COM/2021/347 final, 30 June 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2021:347:FIN.

in 25 countries found that the pieces of primary legislation containing the phrase 'artificial intelligence' grew from one in 2016 to 18 in 2021, many of these specific to a sector or issue.[140] Governments are also considering amendments to existing cross-sectoral regulation such as GDPR, which does not fully anticipate the challenges or the potential of AI.

5.1.2 Impact assessments and audit

The most rapid area of growth concerns algorithmic impact assessments (AIAs) and audits, which attempt to assess and manage ethical risks in the operation of algorithmic systems. While the terminology is not used consistently, AIAs tend to assess impact prospectively (i.e. before a system is in use), while audits are retrospective (i.e. looking back at a period of use).[141]

A number of bodies are currently developing template risk assessments for use by creators or deployers of AI systems. For example, the US National Institute of Standards and Technology (NIST) has released a draft AI Risk Management Framework.[142] The Singapore government is piloting a governance framework and toolkit known as AIVerify.[143] The EU's Artificial Intelligence Act will encourage conformity assessment with technical standards for high-risk AI.[144] The British government is keen to see a new market in AI assurance services established in the UK, by which assurers would certify that AI systems meet their standards and so are trustworthy.[145] The UK's Alan Turing Institute has proposed an assurance framework called HUDERIA.[146] Technical standards bodies are developing frameworks, such as the IEEE's Standard Model Process.[147] There are academic versions, such as capAI,[148] a conformity assessment process designed by a consortium of Oxford-based ethicists, and the European Law Institute's Model Rules on Impact Assessment.[149] There are also fledgling external review processes such as Z-Inspection.[150]

---

140 Stanford University (2022), *Artificial Intelligence Index Report 2022*, https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Chapter-5.pdf, chap. 5.
141 The terminology of 'impact assessment' and 'audit' is used in different ways by different policymakers and academics. For a detailed discussion, see Ada Lovelace Institute and DataKind UK (2020), *Examining the Black Box*, https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf.
142 National Institute of Standards and Technology (2022), *AI Risk Management Framework: Initial Draft*, 17 March 2022, https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf.
143 Infocomm Media Development Authority (2022), *Invitation to Pilot AI Verify AI Governance Testing Framework and Toolkit*, 25 May 2022, https://file.go.gov.sg/aiverify.pdf.
144 McFadden, M., Jones, K., Taylor, E. and Osborn, G. (2021), *Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation*, Oxford Commission on AI & Good Governance, https://oxcaigg.oii.ox.ac.uk/wp-content/uploads/sites/124/2021/12/Harmonising-AI-OXIL.pdf.
145 UK Centre for Data Ethics and Innovation (2021), *The Roadmap to an Effective AI Assurance Ecosystem* https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem.
146 Alan Turing Institute (2021), *Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A proposal prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence*, https://rm.coe.int/huderaf-coe-final-1-2752-6741-5300-v-1/1680a3f688.
147 IEEE Standard Model Process for Addressing Ethical Concerns during System Design, IEEE Std 7000-2021. See also ISO/IEC JTC 1/SC 42 Joint Committee SC 42 on Standardisation in the area of Artificial Intelligence.
148 Floridi, L. et al. (2022), *capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act*, 23 March 2022, http://dx.doi.org/10.2139/ssrn.4064091.
149 European Law Institute (2022), *Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration*, https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Model_Rules_on_Impact_Assessment_of_ADMSs_Used_by_Public_Administration.pdf.
150 Zicari, R. et al. (2021), 'Z-Inspection®: A Process to Assess Trustworthy AI', *IEEE Transactions on Technology and Society*, 2(2), pp. 83–97, https://doi.org/10.1109/TTS.2021.3066209.

Larger businesses have, meanwhile, established their own assessment processes. For example, Google conducts ethical reviews of AI applications it plans to launch.[151] IBM has an AI Ethics Board providing centralized governance, review and decision-making.[152] Rolls-Royce's Aletheia Framework comprises a 32-step practical toolkit for organizations developing and deploying AI.[153]

Typically, AIA processes invite AI developers, providers and users to elicit the ethical values engaged by their systems, refine those values and then assess their proposed or actual AI products and systems (both data and models) against those values, identifying and mitigating risks. Some models take a restrictive view of ethics, focusing primarily on data governance, fairness and procedural aspects rather than all rights.[154] A further tool proposed for data governance is data sheets or 'nutrition labels' that summarize the characteristics and intended uses of data sets, to reduce the risk of inappropriate transfer and use of datasets.[155]

Some governments are introducing impact assessments which are either mandatory or carry strong incentives for compliance. For example, Canada's Directive on Automated Decision-Making requires Canadian government departments to complete and publish an AIA prior to production of any automated decision system.[156] The US's draft Algorithmic Accountability Act, proposed in Congress in 2019 and again in 2022, would require impact assessment of significant automated decisions taken by larger entities.[157] In the UK, the Ada Lovelace Institute has published a detailed proposal for an AIA to be completed by any organization seeking professional access to the National Health Service (NHS)'s proposed National Medical Imaging Platform – the first known AIA for data access in a healthcare context.[158]

While the identification and addressing of ethical risks is a positive step, these processes come with challenges. Risk assessment of AI can mean identifying and mitigating a broad range of impacts on individuals and communities – a task that is potentially difficult, time-consuming and resource-intensive.[159] The identification and mitigation of ethical risks is not straightforward, particularly for teams whose prior expertise may be technical rather than sociological. Extensive engagement with stakeholders may be necessary to obtain a balanced picture of risks. Resourcing challenges are magnified for smaller companies.

---

151 Google (2022), 'AI Principles reviews and operations', https://ai.google/responsibilities/review-process.
152 IBM (2022), 'AI Ethics', https://www.ibm.com/artificial-intelligence/ethics.
153 Rolls Royce (2020), 'The Aletheia Framework', https://www.rolls-royce.com/sustainability/ethics-and-compliance/the-aletheia-framework.aspx.
154 Infocomm Media Development Authority (2022), *Invitation to Pilot AI Verify AI Governance Testing Framework and Toolkit*, 25 May 2022, https://file.go.gov.sg/aiverify.pdf.
155 Gebru, T. et al. (2018), 'Datasheets for datasets', *arXiv*, 1803.09010, https://doi.org/10.48550/arXiv.1803.09010; Data Nutrition Project (2021), 'The Dataset Nutrition Label', https://datanutrition.org/labels (accessed 12 Sep. 2022).
156 Government of Canada (2021), Directive on Automated Decision-Making, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.
157 H.R. 6580 – Algorithmic Accountability Act of 2022, https://www.congress.gov/bill/117th-congress/house-bill/6580/text.
158 Ada Lovelace Institute (2022), *Algorithmic Impact Assessment: A Case Study in Healthcare*, February 2022, https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare.
159 Nonnecke, B. and Dawson, P. (2021), 'Human Rights Implications of Algorithmic Impact Assessments: Priority Considerations to Guide Effective Development', *Carr Center Discussion Paper Series*, Harvard Kennedy School, October 2021, https://carrcenter.hks.harvard.edu/files/cchr/files/nonnecke_and_dawson_human_rights_implications.pdf; Ada Lovelace Institute (2021), *Regulate to innovate*, p.52.

Identification of risks may not even be fully possible before an AI system enters into use, as some risks may only become apparent in the context of its deployment. Hence the importance of ongoing review, as well as review at the design stage. Yet, once a decision has been made to proceed with a technology, many companies have no vocabulary or structure for ongoing discussion of risks. In cases where an AI system is developed by one organization and implemented by another, there may be no system for transferring the initial risk assessment to the recipient organization and for the latter to implement ongoing risk management.

Once risks have been identified, the models offer limited guidance on how to balance competing priorities, including on how to weigh ethical considerations against commercial advantage. Subtle calculations cannot easily be rendered into the simple 'stop' or 'go' recommendation typically required by corporate boards.

Similarly, the audit process presents challenges: auditors may require access to extensive information, including on the operation of algorithms and their impact in context. There is a lack of benchmarks by which to identify or measure factors being audited (such as bias), while audits may not take account of contextual challenges.[160]

British regulators have identified various problems in the current AIA and audit landscape, including a lack of agreed rules and standards; inconsistency of audit focus; lack of access to systems being audited; and insufficient action following audits.[161] There is often inadequate inclusion of stakeholder groups; a lack of external verification; and little connection between these emerging processes and any regulatory regimes or legislation.[162] Recent UK research concluded that public sector policymakers should integrate practices that enable regular policy monitoring and evaluation, including through institutional incentives and binding legal frameworks; clear algorithmic accountability policies and clear scope of algorithmic application; proper public participation and institutional coordination across sectors and levels of governance.[163]

It may be that many algorithms designed without regard to human rights will fail AIAs or audits. As awareness of human rights grows, so much current AI may need adjusting. The Netherlands Court of Audit, having developed an audit framework,[164] recently audited nine algorithms used by the Dutch government. It found that six of those nine failed to meet the requirements of the audit framework on such matters as privacy protection, absence of bias and governance processes.[165]

---

160 Ada Lovelace Institute and DataKind UK (2020), *Examining the Black Box*, p. 10.
161 Digital Regulation Cooperation Forum (2022), *Auditing algorithms: the existing landscape, role of regulators and future outlook*, 28 April 2022, https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook.
162 Ibid., p. 16.
163 Ada Lovelace Institute, AI Now Institute and Open Government Partnership (2021), *Algorithmic Accountability for the Public Sector*, https://www.opengovpartnership.org/wp-content/uploads/2021/08/algorithmic-accountability-public-sector.pdf.
164 Netherlands Court of Audit (2021), 'Understanding Algorithms', 26 January 2021, https://english.rekenkamer.nl/publications/reports/2021/01/26/understanding-algorithms.
165 Netherlands Court of Audit (2022), 'An Audit of 9 Algorithms Used by the Dutch Government', 18 May 2022, https://english.rekenkamer.nl/publications/reports/2022/05/18/an-audit-of-9-algorithms-used-by-the-dutch-government.

Overall, without rigorous implementation of clear standards and external involvement or accountability, there is a risk of 'ethics-washing' rather than genuine mitigation of risks.

5.1.3 Prohibition

Governments and companies are beginning to prohibit forms of AI that raise the most serious ethical concerns. However, there is no consistency in such prohibitions and the rationale behind them is often not openly acknowledged.

For example, some US states have banned certain uses of facial recognition technology, which remain in widespread use in other states. The EU's Artificial Intelligence Act would prohibit certain manipulative AI practices and most use of biometric identification systems in public spaces for law enforcement purposes.[166] Twitter decided to ban political advertising in 2019.[167]

5.1.4 Transparency

A further approach is public transparency measures through registries, release of source code or algorithmic logic (required in France under the Digital Republic Law).[168] In November 2021, the UK government launched the pilot of an algorithmic transparency standard, whereby public sector organizations provide information on their use of algorithmic tools in a standardized format for publication online. Several government algorithms have since been made public as a result.[169]

5.1.5 Procurement conditions

There is likely to be a rapid growth in the imposition of conditions in the sale of algorithmic systems, particularly where purchasers such as governments and local authorities will be seeking to use those systems in the public interest. Authorities are likely to impose contractual conditions requiring the system to respect stipulated criteria on such matters as bias and transparency. For example, the City of Amsterdam has developed contractual terms requiring suppliers of AI and algorithmic systems to meet standards of explainability and transparency, including on what data is used and how bias is counteracted.[170] Such conditions imposed by the public sector may have the effect of driving up standards more widely.

---

166 Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, Article 5.
167 Twitter (2019), 'Political Content', https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html.
168 Loi No. 2016-1321 du 7 octobre 2016 pour une République Numerique.
169 Central Digital and Data Office (2021), 'Algorithmic Transparency Standard', https://www.gov.uk/government/collections/algorithmic-transparency-standard.
170 Gemeente Amsterdam (2022), 'Contractual terms for algorithms', https://www.amsterdam.nl/innovatie/digitalisering-technologie/algoritmen-ai/contractual-terms-for-algorithms.

5.2 Processes: human rights law

5.2.1 Governmental duty to protect against breaches

Governments have a duty both to comply with human rights in any uses of AI they adopt – for example, in public decision-making – and to protect individuals from abuses of human rights by companies and other non-state actors. States must take 'appropriate steps to prevent, investigate, punish and redress such abuse through effective policies, legislation, regulations and adjudication'.[171]

Governments are expected to find the appropriate mix of laws, policies and incentives to protect against human rights harms. A 'smart mix' of national and international, mandatory and voluntary measures would help to foster business respect for human rights.[172] This includes requiring companies to have suitable corporate structures to identify and address human rights risk on an ongoing basis, and to engage appropriately with external stakeholders as part of their human rights assessments. Where businesses are state-owned, or work closely with the public sector, the government should take additional steps to protect against human rights abuses through management or contractual control.[173]

Governments' human rights obligations mean that they cannot simply wait and see how AI develops before engaging in governance activities. They are obliged to take action, including via regulation and/or the imposition of impact assessments and audits, to ensure that AI does not infringe human rights. Governments should ensure that they understand the implications of human rights for AI governance, deploying a dedicated capacity-building effort or technology and human rights office where a gap exists.[174]

There is an urgent need for governments to devise regulation that is both effective in ensuring that companies do not infringe individuals' human rights when designing and implementing AI systems, and that provides for effective remedies in the event of any such infringement. Given the ambiguity of commitments to ethics and the strength of countervailing commercial considerations, a purely voluntary approach is unlikely to protect individuals' human rights adequately. Indeed, some argue that states are obliged to enact legally binding norms to protect human rights in light of the challenges posed by AI systems.[175] Governments should regulate to either prohibit or require constraints on applications of AI, such as biometric technologies, that risk interfering with human rights in a manner clearly disproportionate to any countervailing legitimate interest.

---

171 UN Office of the High Commissioner for Human Rights (2011), *Guiding Principles on Business and Human Rights*, principle 1.
172 Ibid., principle 3; and UN OHCHR B-Tech (2021), *Bridging Governance Gaps in the Age of Technology – Key characteristics of the State Duty to Protect*, https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/b-tech-foundational-paper-state-duty-to-protect.pdf.
173 UN Office of the High Commissioner for Human Rights (2011), *Guiding Principles on Business and Human Rights,* principle 4; UN OHCHR B-Tech (2021), *Bridging Governance Gaps in the Age of Technology – Key characteristics of the State Duty to Protect*.
174 Element AI (2019), *Closing the Human Rights Gap in AI Governance*, http://mediaethics.ca/wp-content/uploads/2019/11/closing-the-human-rights-gap-in-ai-governance_whitepaper.pdf.
175 Bello y Villarino, J.-M. and Vijeyarasa, R. (2022), 'International Human Rights, Artificial Intelligence, and the Challenge for the Pondering State: Time to Regulate?', *Nordic Journal of Human Rights*, 40(1), pp. 194–215, https://doi.org/10.1080/18918131.2022.2069919.

Governments should ensure that AIA and audit processes are conducted systematically, employing rigorous standards and due process, and that such processes pay due regard to potential human rights impacts of AI: for example by making assessment of human rights risks an explicit feature of such processes.[176] To incentivize corporate good practice, demonstrate respect for human rights and facilitate remedy, states should also consider requiring companies to report publicly on any due diligence undertaken and on human rights impacts identified and addressed.

Supervision by regulatory and administrative authorities is an important element of accountability for compliance with human rights responsibilities, in parallel with legal liability for harms.

Supervision by regulatory and administrative authorities is an important element of accountability for compliance with human rights responsibilities, in parallel with legal liability for harms. As some European countries and the EU begin to implement mandatory human rights and environmental due diligence obligations for larger businesses,[177] human rights experts are exploring administrative supervision of corporate duties as a complement to liability for harms in the courts.[178]

Governments have legal obligations not to breach human rights in their provision of AI-assisted systems. Anyone involved in government procurement of AI should have enough knowledge and information to understand the capacity and potential implications of the technology they are buying, and to satisfy themselves that it meets required standards on equality, privacy and other rights (such as the Public Sector Equality Duty in the UK). Governments should negotiate the terms of public–private contracts and deploy procurement conditions to ensure that AI from private providers is implemented consistently with human rights. They should also take steps to satisfy themselves that this requirement is met. Public procurement is a means of encouraging improvements to human rights standards in the AI industry as a whole.[179] It is important also to ensure that AI systems already adopted comply with human rights standards: the experience of the Netherlands demonstrates that systems adopted to date can be problematic.[180]

---

176 Nonnecke, B. and Dawson, P. (2022), *Human Rights Impact Assessments for AI: Analysis and Recommendations*, New York: Access Now, October 2022, https://www.accessnow.org/cms/assets/uploads/2022/11/Access-Now-Version-Human-Rights-Implications-of-Algorithmic-Impact-Assessments_-Priority-Recommendations-to-Guide-Effective-Development-and-Use.pdf.
177 European Commission (2022), 'Proposal for a Directive on Corporate Sustainability Due Diligence', 23 February 2022, https://ec.europa.eu/info/publications/proposal-directive-corporate-sustainable-due-diligence-and-annex_en. Several EU member states and other states have implemented similar obligations or elements of mandatory human rights due diligence. For example, see Office of the UN High Commissioner for Human Rights (2020), *UN Human Rights "Issues Paper" on legislative proposals for mandatory human rights due diligence by companies*, June 2020, https://www.ohchr.org/sites/default/files/Documents/Issues/Business/MandatoryHR_Due_Diligence_Issues_Paper.pdf, pp. 3–5.
178 Shift and Office of the UN High Commissioner for Human Rights (2021), *Enforcement of Mandatory Due Diligence: Key Design Considerations for Administrative Supervision*, Policy Paper, October 2021, https://shiftproject.org/wp-content/uploads/2021/10/Enforcement-of-Mandatory-Due-Diligence_Shift_UN-Human-Rights_Policy-Paper-2.pdf.
179 Office of the UN High Commissioner for Human Rights (2022), *The Practical Application of the Guiding Principles on Business and Human Rights to the Activities of the Technology Sector*, April 2022, https://reliefweb.int/report/world/practical-application-guiding-principles-business-and-human-rights-activities-technology-companies-report-office-united-nations-high-commissioner-human-rights-ahrc5056-enarruzh, para. 20.
180 Netherlands Court of Audit (2022), 'An Audit of 9 Algorithms Used by the Dutch Government'.

5.2.2 Corporate responsibility to respect human rights

The UN's Guiding Principles on Business and Human Rights are clear that 'business enterprises should respect human rights'. In other words, companies (particularly large ones)[181] should avoid infringing human rights and should address any adverse human rights impacts resulting from their activities.[182] Companies should have a policy commitment to meet their human rights responsibilities, approved at senior level, publicly available and embedded in the culture of the business.[183] Companies must also have an ongoing due diligence process of human rights impact assessment, tracked for responsiveness and reported externally, which allows them to identify, mitigate and remedy human rights impacts.[184] By deploying a responsible business agenda, identifying and mitigating risks, companies can forestall problems and save themselves the time, money and acrimony of litigation.

Due diligence in the AI context is particularly challenging because of two distinguishing features. First, AI's capacity for self-improvement may make it difficult to predict its consequences. Second, AI's human rights impact will depend not only on the technology itself, but also on the context in which it is deployed. In light of both these factors, due diligence on AI applications that may affect human rights must be extensive and involve as wide a set of stakeholders as may be affected by the AI. Further, given the risk of unanticipated consequences, AI must be reviewed regularly once in operation. Hence, the former UN high commissioner on human rights called for comprehensive human rights due diligence to be conducted 'when AI systems are acquired, developed, deployed and operated',[185] with that due diligence to continue 'throughout the entire life cycle of an AI system'[186] and to include consultations with stakeholders and involvement of experts.[187] At present, many companies lack structures and processes to detect and act on human rights issues on an ongoing basis. The former UN high commissioner also called for the results of due diligence to be made public.[188]

Some companies' AIAs are labelled as human rights assessment, like Verizon's ongoing human rights due diligence.[189] Other AI ethics assessments, such as that adopted by the IEEE and the proposed AIA for the National Medical Imaging Platform, look similar to human rights due diligence, but are not labelled as such. Google reviews proposals for new AI deployment by reference to its AI Principles, a process that can include consultation with human rights experts.[190]

---

181 The UN *Guiding Principles on Business and Human Rights* apply to all businesses, but the extent of business responsibilities increases with the organization's size and the impact of its work: see UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights (2011)*, principle 14.
182 UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights*, principles 11, 13.
183 UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights*, principles 15 and 16.
184 UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights*, principles 15–21. See also Data & Society and European Center for Non-Profit Law (2021), *Recommendations for Assessing AI Impacts to Human Rights, Democracy and the Rule of Law*, https://ecnl.org/sites/default/files/2021-11/HUDERIA%20paper%20ECNL%20and%20DataSociety.pdf.
185 United Nations High Commissioner for Human Rights (2021), *The Right to Privacy in the Digital Age*, A/HRC/48/31, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/249/21/PDF/G2124921.pdf, para. 48.
186 Ibid., para. 49.
187 Ibid., para. 50.
188 Ibid., para. 50.
189 Verizon (2022), 'Human Rights at Verizon', https://www.verizon.com/about/investors/human-rights-at-verizon.
190 Google AI (2022), 'AI Principles reviews and operations'. https://ai.google/responsibilities/review-process.

Whatever the labelling, certain features of human rights impact assessment are commonly omitted from corporate processes:

- **Transparency.** General statements of corporate intention and activity are easier to find than public statements of human rights risks actually identified and mitigated through due diligence processes.
- **Scope.** Some corporate processes only cover specific issues, such as bias and privacy, rather than the full range of human rights, or make only brief mention of other rights.[191]
- **Effect.** It is often not clear what effect impact assessments have on the company's activities.[192] Human rights due diligence requires that human rights risks be mitigated, whereas some business processes seem to entail balancing risks against perceived benefits.[193]
- **Duration.** Human rights due diligence includes a requirement for ongoing review post-implementation, whereas many corporate reviews appear to focus only on product development. Ongoing review is particularly important in light of AI's capacity for self-improvement over time. Otherwise, there is a risk that assessments give algorithmic processes a veneer of legitimacy rather than genuinely having an impact on activities.[194] This risk is amplified when there is no transparency about the process, its results or impact.

In addition to ensuring the adequacy of their impact assessment processes from a human rights perspective, companies should foster a pro-human rights culture throughout their organization. This means ensuring that AI teams are representative of society's diversity and the diversity of intended consumers, such that equality is 'baked in' to system design. It means engaging adequate internal and external expertise to conduct human rights due diligence and impact assessments, including through involvement of stakeholders, and commitment at board level to addressing human rights impacts identified. It also means public reporting of any human rights risks and impacts identified and measures taken. It may mean providing training on human rights for all those working on AI – including technical experts, engineers and devisers of technical standards. It must include ongoing monitoring of human rights impacts over time and preparedness to address new concerns that may arise.

---

191 For example, Microsoft's *Responsible AI Impact Template* (2022), https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf.
192 Ada Lovelace Institute and DataKind UK (2020), *Examining the Black Box*, p. 18.
193 For example, Google states that it will not pursue technologies that cause or are likely to cause overall harm, and 'where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.' They also say that they will not proceed with 'technologies whose purpose contravenes widely accepted principles of international law and human rights.' See Google AI (2022), 'Artificial Intelligence at Google: Our Principles', https://ai.google/principles.
194 Ada Lovelace Institute and DataKind UK give the example of an impact evaluation of a predictive risk modelling tool for Allegheny County, PA's children's welfare office, with positive results that both conflicted with other reviews of the tool and may have provided legitimacy for further use of AI in children's social services. Ada Lovelace Institute and DataKind UK (2020), *Examining the Black Box,* p. 19.

06 Remedies in AI governance: the contribution of human rights

Both governments and companies should provide suitable access to remedy for when AI goes wrong. This entails effective reparation, accountability and measures to prevent recurrences.

6.1 Remedies: the landscape

Little attention has been given to the development of a scheme of remedies for when AI goes wrong. Responsibility needs to be clarified, and transparency is required to assess whether and how AI has gone wrong.

While AI governance principles commonly include a principle of accountability, this often refers to impact assessments, audit or oversight, rather than a requirement of remedy in the event of harms.[195] Many sets of AI governance principles in fact have no provision for remedy. As the UN special rapporteur on contemporary forms of racism has pointed out, '[e]thical commitments have little measurable

195 For example, UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021) discusses oversight, impact assessment, audit and due diligence mechanisms (paras 42 and 43) and suggests that states may wish to consider establishing an ethics commission or ethics observatory (para. 133).

effect on software development practices if they are not directly tied to structures of accountability in the workplace'.[196]

To some extent, legal remedies for wrongs caused by the application of AI already exist in tort law (negligence) and administrative law, particularly where those wrongs are on the part of public authorities. However, the law and its processes will need to develop metrics for evaluating AI. For example, English administrative law typically has regard to whether the decision-maker took the right factors into account when making their decision. But AI relies on statistical inferences rather than reasoning. Factors such as the opacity of AI systems and imbalance of information and knowledge between companies and users, scalability of errors and rigidity of decision-making may also pose challenges.[197] As yet, there is no clear 'remedy pathway' for those who suffer abuses of human rights as a result of the operation of AI.[198]

Those at greatest risk from harms caused by AI are likely to be the most marginalized and vulnerable groups in society, such as immigrants and those in the criminal justice system.

Those at greatest risk from harms caused by AI are likely to be the most marginalized and vulnerable groups in society, such as immigrants and those in the criminal justice system. This makes it all the more important to ensure that avenues for remedy are accessible to all, whatever their situation.

There has already been some litigation challenging the application of AI by reference to human rights law or its local equivalent. Notable cases include:

- In 2016, *State of Wisconsin v Eric L Loomis*, which challenged the use of AI COMPAS risk assessments when sentencing defendants in criminal cases. The COMPAS risk assessment was an assessment of recidivism risk, based on comparisons with other individuals with a similar history of offending. The Supreme Court of Wisconsin held that a court's consideration of a COMPAS risk assessment is consistent with the defendant's right to due process, provided that the risk assessment is used in parallel with other factors and is not determinative of the defendant's sentence.[199]
- In May 2017, teachers in Houston successfully challenged the use of an algorithm known as EVAAS,[200] developed by a private company to measure teacher effectiveness.[201] The aim of the algorithm was to enable the Houston

196 Report of the UN Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, *Racial discrimination and emerging digital technologies: a human rights analysis,* A/HRC/44/57 18 June 2020, para. 62.
197 See Williams, R. (2021), 'Rethinking Administrative Law for Algorithmic Decision Making', *Oxford Journal of Legal Studies*, 2(2), https://doi.org/10.1093/ojls/gqab032, pp. 468–94.
198 Office of the UN High Commissioner for Human Rights (2022), *The Practical Application of the Guiding Principles on Business and Human Rights to the Activities of the Technology Sector*, para. 58.
199 *State of Wisconsin v Eric L Loomis* 2016 WI 68, 881 N.W.2d 749.
200 EVAAS stands for Educational Value-Added Assessment System.
201 *Houston Federation of Teachers v Houston Independent School District* 251 F.Supp.3d 1168 (SD Tex 2017).

Independent School District (HISD) to terminate the employment of teachers whose performance was deemed ineffective. The US district court denied HISD's application for summary judgment against the teachers' claim. The court found that the teachers were 'unfairly subject to mistaken deprivation of constitutionally protected property interests in their jobs', contrary to the Due Process Clause of the Fourteenth Amendment of the US Constitution, because they had no meaningful way to ensure correct calculation of their scores, nor opportunity to independently verify or replicate those scores. After the summary judgment, the case was settled and HISD abandoned the EVAAS system.[202]

- In March 2018, Finland's National Non-Discrimination and Equality Tribunal decided that a credit institution's decision not to grant credit to an individual was discriminatory. The tribunal ruled that the credit institution's decision was made not on the basis of the individual's own credit behaviour and creditworthiness, but by drawing assumptions from statistical data and information on payment default relating to other people, by criteria such as gender, first language, age and residential area. The tribunal prohibited the credit institution from using this decision-making method.[203]
- In February 2020, the Hague district court ordered the Dutch government to cease its use of SyRI, an automated programme that reviewed the personal data of social security claimants to predict how likely people were to commit benefit or tax fraud. The Dutch government refused to reveal how SyRI used personal data, such that it was extremely difficult for individuals to challenge the government's decisions to investigate them for fraud or the risk scores stored on file about them. The Court found that the legislation regulating SyRI did not comply with the right to respect for private life in Article 8 ECHR, as it failed to balance adequately the benefits SyRI brought to society with the necessary violation of private life caused to those whose personal data it assessed. The Court also found that the system was discriminatory, as SyRI was only used in so-called 'problem neighbourhoods', a proxy for discrimination on the basis of socio-economic background and immigration status.[204]
- In August 2020, *R (Bridges) v Chief Constable of South Wales Police*[205] was the first challenge to AI invoking UK human rights law. South Wales Police was trialling the use of live automated facial recognition technology (AFR) to compare CCTV images of people attending public events with images of persons on a database. If there was no match, the CCTV images were immediately deleted from the AFR system. The complainant challenged AFR's momentary capture of his image and comparison with its watch-list database, by reference to Article 8 ECHR and the UK Data Protection Act. The Court

202 McCully, J. (2017), 'Houston Federation of Teachers and Others v HISD', Atlas Lab blog, https://www.atlaslab.org/post/houston-federation-of-teachers-and-others-v-hisd-secret-algorithm-used-to-fire-teachers.
203 National Non-Discrimination and Equality Tribunal of Finland (2018), *Assessment of creditworthiness, authority, direct multiple discrimination, gender, language, age, place of residence, financial reasons, conditional fine*, Register No. 216/2017, 21 March 2018, https://www.yvtltk.fi/material/attachments/ytaltk/tapausselosteet/45LI2c6dD/YVTltk-tapausseloste-_21.3.2018-luotto-moniperusteinen_syrjinta-S-en_2.pdf.
204 Toh, A. (2020), 'Dutch Ruling a Victory for Rights of the Poor', *Human Rights Watch Dispatches*, 6 February 2020, https://www.hrw.org/news/2020/02/06/dutch-ruling-victory-rights-poor.
205 [2020] EWCA Civ 1058.

of Appeal found that there was not a proper basis in law for the use of AFR. Consequently, its use breached the Data Protection Act. The court declined to find that the police's use of AFR struck the wrong balance between the rights of the individual and the interests of the community. But it did find that South Wales Police had failed to discharge the statutory Public Sector Equality Duty,[206] because in buying the AFR software from a private company and deploying it, they had failed to take all reasonable steps to satisfy themselves that the software did not have a racial or gender bias (notwithstanding that there was no evidence to support the contention that the software was biased). The case therefore temporarily halted South Wales Police's use of facial recognition technology, but allowed the possibility of its reintroduction in future with proper legal footing and due regard to the Public Sector Equality Duty. Indeed, South Wales Police has since reintroduced facial recognition technology for use in certain circumstances.[207]

- The Italian courts, having held in 2019 that administrative decisions based on algorithms are illegitimate, reversed that view in 2021. The courts welcomed the speed and efficiency of algorithmic decision-making but clarified that it is subject to general principles of administrative review in Italian law, including transparency, effectiveness, proportionality, rationality and non-discrimination. Complainants about public decision-making are entitled to call for disclosure of algorithms and related source code in order to challenge decisions effectively.[208]
- In July 2022, the UK NGO Big Brother Watch issued a legal complaint to the British information commissioner in respect of alleged use of facial recognition technology by Facewatch and the supermarket chain Southern Co-op to scan, maintain and assess profiles of all supermarket visitors in breach of data protection and privacy rights.[209]

6.2 Remedies: human rights law

Human rights law requires both governments and companies to provide a suitable right to remedy in the event of breach of their obligations and responsibilities.[210] Remedy comprises effective reparation, appropriate accountability for those responsible, as well as measures to prevent recurrences. The availability of remedy is crucial if human rights or ethical principles are to have real impact in the face of countervailing commercial considerations.

---

206 Section 149(1) Equality Act 2010: 'A public authority must, in the exercise of its functions, have due regard to the need to- (a) eliminate discrimination, harassment, victimisation and any other conduct that is prohibited by or under this Act; (b) advance equality of opportunity between persons who share a relevant protected characteristic and persons who do not share it; (c) foster good relations between persons who share a relevant protected characteristic and persons who do not share it.'
207 South Wales Police (2022), 'Facial Recognition Technology', https://www.south-wales.police.uk/police-forces/south-wales-police/areas/about-us/about-us/facial-recognition-technology.
208 Liguori, L. and Vittoria La Rosa, M. (2021), 'Law and Policy of the Media in a Comparative Perspective', *Filodiritto* blog, 20 May 2021, https://www.filodiritto.com/law-and-policy-media-comparative-perspective.
209 Big Brother Watch (2022), *Grounds of Complaint to the Information Commissioner under section 165 of the Data Protection Act 2018: Live Automated Facial Recognition by Facewatch Ltd and the Southern Cooperative Ltd*, https://docs.reclaimthenet.org/big-brother-watch-co-op-facewatch-legal-complaint.pdf.
210 International Covenant on Civil and Political Rights, Article 2(3); UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights*, principles 25–31.

This means that, at all stages of design and deployment of AI, it must be clear who bears responsibility for its operation. In particular, clarity is required on where the division of responsibilities lies between the developer of an AI system and the purchaser and deployer of the system, including if the purchaser adapts the AI or uses it in a way for which it was not intended. Consequently, purchasers of AI systems will need adequate understanding or assurance as to how those systems work, as was demonstrated for the public sector in the *Bridges* case, discussed above. In that case, the court also held that commercial confidentiality around any AI technology does not defeat or reduce the requirement for compliance with the Public Sector Equality Duty.[211]

Complainants need to know how to complain and to whom, and to be confident that their complaint will be addressed in a timely manner. Remedy relies on transparency and explainability – complainants should have enough information to understand how a decision about them was made, and the role and operation of AI in the decision-making process. They may need access to data on how the AI was designed and tested, how it was intended to operate and how it has operated in the specific case, as well as information on the role of human decision-making or oversight in the process.

Remedy may be provided by the courts, by other governmental mechanisms such as regulators, ombudspersons and complaints processes, as well as by non-governmental mechanisms such as corporate remediation processes. The UN Guiding Principles recommend that all businesses 'establish or participate in effective operational-level grievance mechanisms'.[212] Such mechanisms should be legitimate (i.e. enabling trust); accessible; predictable; equitable; transparent; rights-compatible; a source of continuous learning; and based on engagement and dialogue with stakeholders.[213]

There are challenges in designing appropriate grievance mechanisms for addressing harms caused by AI. Remedial systems that rely on individual complaint tend to be better at addressing significant harms suffered by few than harms suffered by many.[214] But AI, with its capacity for operation at scale, risks infringing the rights of large numbers of people – for example, by using personal data in violation of the right to privacy or engaging in widespread discriminatory treatment. Many of the people affected could be vulnerable or marginalized, including asylum-seekers and those in the criminal justice system. Consequently, there needs to be provision both for individual complaints and for group or representative complaints against a whole system rather than a single decision. Ombudsmen, national human rights institutions and civil society organizations should be adequately equipped to support victims' complaints and to challenge AI systems that are systematically causing harm. Remedies should consist both of adequate remedy to victims and requirements to improve, or end the use of, AI systems to prevent recurrence of any harm identified.

---

211 *R (Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058, para. 199.
212 UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights*, principle 29.
213 UN Office of the High Commissioner on Human Rights (2011), *Guiding Principles on Business and Human Rights*, principle 31.
214 Raso, F. et al. (2018), *Artificial Intelligence & Human Rights: Opportunities & Risks*, Berkman Klein Center Research Publication No. 2018-6, 25 September 2018, http://dx.doi.org/10.2139/ssrn.3259344, p. 56.

Similarly, a business should be able to pursue accountability against other companies that have harmed its operations as a result of AI. This may be because the business has purchased an AI system that has not functioned as intended, or because another company's AI has in some way interfered with its operations.

Many challenges are expected in this field in the coming years. The guiding principle should remain provision of an effective right to remedy, including for breach of human rights responsibilities.

07 Conclusion and recommendations

To place human rights at the heart of AI governance, companies, governments, international organizations, civil society and investors must take effective practical steps.

As AI begins to reshape the human experience, human rights must be central to its governance. There is nothing to fear, and much to gain, from taking human rights as the baseline for AI governance.

Failure to take account of human rights means setting aside well-established, widely acknowledged parameters of liberty, fairness and equality, as well as processes and accountability for their implementation. It involves creating confusing and inadequate alternatives to existing norms. It also duplicates much of the work of developing those norms, the processes for their implementation and the remedies for their breach.

If human rights are to be placed at the centre of AI governance, the following practical actions are necessary.

For companies:

- Continue to promote AI ethics and responsible business agendas, while acknowledging the important complementary role of existing human rights frameworks;
- Champion a holistic commitment to all human rights standards from the top of the organization. Enable a change of corporate mindset, such that human rights are seen as a useful tool in the box rather than as a constraint on innovation;
- Recruit people with human rights expertise to join AI ethics teams to encourage multi-disciplinary thinking and spread awareness of human rights organization-wide. Use human rights as the common language and framework for multi-disciplinary teams addressing aspects of AI governance;

- Conduct human rights due diligence and adopt a human rights-based approach to AI ethics and impact assessment. Create decision-making structures that allow human rights risks to be monitored, flagged and acted upon on an ongoing basis;
- Ensure uses of AI are explainable and transparent, so that people affected can find out how an AI or AI-assisted decision was, or will be, made; and
- Establish a mechanism for individuals to seek remedy if they are dissatisfied with the outcome of a decision made or informed by AI.

For governments:

- Ensure adequate understanding of human rights among government officials and place human rights at the heart of AI regulation and policies, either via the establishment of a dedicated office or other existing mechanisms;
- Equip teams involved in government procurement of systems and services with expertise in AI and human rights. Use contracting policy and procurement conditions to increase compliance with human rights standards among businesses;
- Establish a discussion forum on AI governance that engages all stakeholders, including human rights advocates, to foster better understanding and mutual benefit from others' perspectives;
- Ensure that technical standards bodies, AI assurance mechanisms and devisers of algorithmic impact assessment and audit processes give due regard to human rights when developing and monitoring standards for AI governance;
- Consider cross-cutting regulation to ensure that AI deployed by both the public and private sectors meets human rights standards;
- Put in place human rights-compatible standards and oversight for AIAs and audits, as well as adequate provision of remedy for alleged breaches;
- Educate the public on the vital role of human rights in protecting individual freedoms as AI technology develops. Offer guidance to schools and teachers so that children have an understanding of human rights before they encounter AI;
- Ensure that all uses of AI are explainable and transparent, such that people affected can find out how an AI or AI-informed decision was, or will be, made;
- Provide adequate resources for national human rights bodies and regulators, such as the UK Equalities and Human Rights Commission, to champion the role of human rights in AI governance. Ensure these bodies are included in discussions on emerging tech issues;
- Incentivize AI development that benefits society as widely as possible and contributes to implementation of the UN's SDGs; and
- Liaise with other governments and international organizations with a view to harmonizing understanding of the impact of international human rights law on the development and implementation of AI (for example, through use of soft law and guidance).

For the UN and other international/regional organizations:

- Adopt consensus principles on AI and human rights that clarify the duties of states and responsibilities of companies in this field, as well as the requirements for remedy. Publish a sister document to the UN's Guiding Principles on Business and Human Rights to outline these principles, accessible to all stakeholders including software developers and engineers;
- Establish a new multi-stakeholder forum that brings together the tech and human rights communities, as well as technical standards bodies, to discuss challenges around the interaction of human rights and technology, including AI.[215] A regular, institutionalized dialogue would raise levels of understanding and cooperation on all sides of the debate, and would help prevent business exploitation of legal grey areas;[216]
- Ensure, via the UN secretary-general's envoy on technology, that all parts of the UN (including technical standards bodies and procurement offices) align with the OHCHR in placing human rights at the centre of their work on technology;
- Continue to promote UNESCO's Recommendation on the Ethics of Artificial Intelligence, including the international human rights obligations and commitments to which it refers, facilitating knowledge-sharing and capacity-building to enable effective implementation in all states;
- Advance dialogue and coherent approaches to the implications of AI for human rights, via treaties or soft law, and support national governments in their governance of AI;
- Conduct human rights due diligence before deploying AI; and
- Integrate AI into development and capacity-building activities to accelerate implementation of the SDGs.

For civil society and academics:

- Push for inclusion in the AI governance conversation, including by fostering connections with the software development community and corporate public policy teams;
- Debunk human rights myths. Explain to a wide array of audiences (including business leaders, investors and governments) that human rights are reasonable not radical; that human rights do not stymie innovation but establish a level playing field in guarding against egregious development.
- Demonstrate the positive role of human rights as a regulatory system by reference to existing processes of human rights due diligence and remedy;

---

215 As discussed at the Digital Democracy Dialogue 3D2, in Montreux, Switzerland, in November 2021. See also Universal Rights Group (2021), *Placing Digital Technology at the Service of Democracy and Human Rights*, https://www.universal-rights.org/wp-content/uploads/2021/12/3D2_designed-report_V1.pdf.
216 Human Rights Council Advisory Committee (2021), *Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights*, A/HRC/47/52, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/110/34/PDF/G2111034.pdf?OpenElement, para. 55.

- Encourage inter-disciplinary engagement at universities and raise awareness of human rights in technology-focused studies – for example, by introducing human rights as an element of computer science degrees and coding 'bootcamps';
- Facilitate collaboration between civil society and the software development community on the development and use of AI to achieve of the SDGs; and
- Test the implications of human rights for AI through strategic litigation.

For investors:

- Include assessment of the implications of AI for human rights in ESG or equivalent investment metrics.[217]

217 Minkkinen, M., Niukkanen, A., and Mäntymäki, M. (2022), 'What about investors? ESG analyses as tools for ethics-based AI auditing', *AI & Society*, https://doi.org/10.1007/s00146-022-01415-0.

About the author

Kate Jones is an associate fellow with Chatham House's International Law Programme. She is a consultant and researcher on human rights law, public international law, governance and diplomacy, focusing on their intersection with technology. With more than 20 years of legal experience, Kate has published and spoken widely on aspects of tech governance, with specialisms in disinformation and foreign interference, artificial intelligence and human rights. Her publications at Chatham House include the 2019 research paper *Online Disinformation and Political Discourse: Applying a Human Rights Framework*.

Acknowledgments

This research paper is published as part of the Human Rights Pathways initiative of Chatham House, funded by the Swiss Federal Department of Foreign Affairs.

Thanks are due to all those whose ideas and comments have helped shape the paper. This includes those who generously agreed to be interviewed; all the participants in a Chatham House roundtable on Artificial Intelligence and Human Rights, convened with kind cooperation of the Geneva Human Rights Platform at the Villa Moynier, Geneva in May 2022; and all who attended a London meeting on AI and human rights kindly convened by the European Center for Not-for-Profit Law (ECNL) in June 2022.

The author is grateful to all at Chatham House who have contributed to the content, editing and publication of the paper, including Harriet Moynihan, Chanu Peiris, Rashmin Sagoo, Elizabeth Wilmshurst KC, Marjorie Buchser, David Griffiths, Rowan Wilkinson, Rachel Mullally, Sophia Rose and Chris Matthews. She would also like to thank those outside Chatham House who reviewed drafts: Vanja Škorić of ECNL, Janis Wong of the Alan Turing Institute and the anonymous peer reviewers.

Finally, thanks go to many others for interesting conversations and debates on this topic over the last year, including Lukas Madl and the advisory board of Innovethic, Christian Hunt and his Human Risk Podcast, and the Digital Society Initiative at Chatham House.

Cover image: An attendee tries a virtual reality experience during the Mobile World Congress trade show in Barcelona, Spain on 3 March 2022.
Photo credit: Copyright © Joan Cros/NurPhoto/Getty Images

ISBN 978 1 78413 549 2
DOI 10.55317/9781784135492

Cite this paper: Jones, K. (2023), *AI governance and human rights: Resetting the relationship,* Research Paper, London: Royal Institute of International Affairs, https://doi.org/10.55317/9781784135492.

This publication is printed on FSC-certified paper.
designbysoapbox.com

Independent thinking since 1920

The Royal Institute of International Affairs
Chatham House
10 St James's Square, London SW1Y 4LE
T +44 (0)20 7957 5700
contact@chathamhouse.org | chathamhouse.org

Charity Registration Number: 208223