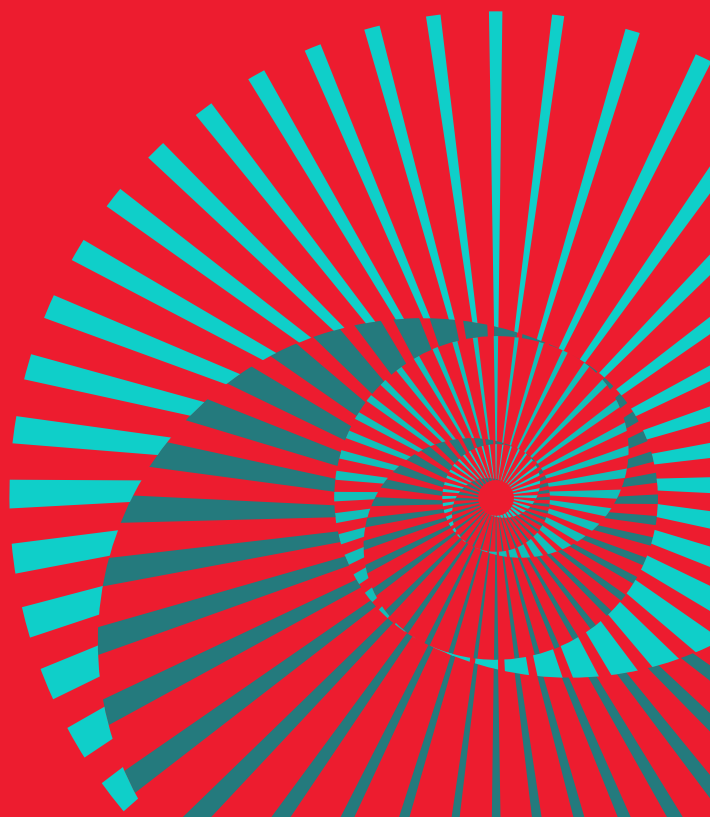
Think Again: Freedom of Thought in the Age of Artificial Intelligence

POLICY BRIEF



OSCE ODIHR

Think Again: Freedom of Thought in the Age of Artificial Intelligence

POLICY BRIEF

Acknowledgements

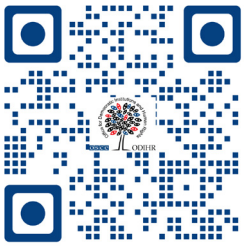
ODIHR extends its sincere gratitude to those who generously shared their knowledge and expertise in the preparation of this document. Their thoughtful reviews and feedback at various stages of the process have been invaluable. ODIHR thanks the members of the ODIHR Panel of Experts on Freedom of Religion or Belief, with special appreciation to Ahmed Shaheed, David Griffiths and Pasquale Annicchino for their insightful input. Thanks also go to the following for reviewing various drafts and offering important contributions: Cameran Ashraf, Christoph Bublitz, Hervé Chneiweiss, Emily Elstub, Alexander Kriebitz and Simon McCarthy-Jones, as well as to colleagues within the office of the Representative of Freedom of the Media. Special thanks go to Benjamin Greenacre, whose discussions on this topic greatly enriched the final outcome and to Milena Costas Trascasas for her continued dialogue and spirit of cooperation.

Think Again: Freedom of Thought in the Age of Artificial Intelligence

Published by the OSCE Office for Democratic Institutions and Human Rights (ODIHR)
ul. Miodowa 10
00-251 Warsaw
Poland
www.osce.org/odihr

ISBN: 978-92-9271-453-6

Designed by Filip Andronik



Contents

Executive summary

This policy brief provides a succinct introduction to the right to freedom of thought, noting that states have a positive duty to create an enabling environment for its full realization.[1] It begins by framing freedom of thought as a foundational right that must be protected and promoted in all contexts. It then explores how emerging technologies, particularly artificial intelligence (AI) and neurotechnologies, are reshaping the landscape of thought. While such technologies, depending on how they are developed and deployed, can be used to enhance the enjoyment of human rights[2] — for example, by improving access to diverse information, tailored to aid educational processes or support for medical advancements — they also present unprecedented risks. These require comprehensive assessments and robust safeguards before development, deployment or use.

The brief examines the implications for freedom of thought of the new information ecosystem, as well as the potential impacts of AI-based neurotechnologies, which raise fundamental questions about the inviolability of individual mental autonomy. The policy brief also outlines recommendations for states to consider in upholding freedom of thought in the digital age. These include:

- Regulatory and institutional safeguards to address emerging risks, particularly those posed by AI and neurotechnologies
- Ethical oversight mechanisms and mandatory human rights impact assessments embedded across the AI life cycle[3]
- Initiatives to ensure transparency, inclusive public debate and accountability in decision-making processes

1 While recognizing the significant implications for freedom of thought of certain AI applications in the military, or in surveillance contexts, these areas are beyond the scope of this brief.

2 Theresa Addie, MS., Harnessing Technology to Safeguard Human Rights: AI, Big Data, and Accountability, Human Rights Research Centre, 8 April 2025; UN Special Rapporteur (UNSR) on Freedom of Religion or Belief (FoRB) Ahmed Shaheed, Report of the Special Rapporteur on freedom of religion or belief, A/76/380, 5 October 2021, paras. 2-4.

3 The UN Global Digital Compact, annexed to United Nations General Assembly resolution A/RES/79/1, on the Pact for the Future, adopted on 22 September 2024, includes the following under the life cycle of digital and emerging technologies: pre-design, design, development, evaluation, testing, deployment, use, sale, procurement, operation and decommissioning stages.

- Capacity-building through education and training on ethics and human rights for developers and operators of AI technologies
- Strengthened roles for independent institutions in protecting freedom of thought and democratic values

The brief concludes that innovation that fails to consider and protect human rights from the outset will undermine the goals that should guide ethical, scientific and technological progress, namely to benefit humankind.[4]

Building on work by international organizations and academia on AI, neurotechnologies and human rights, this policy brief highlights the specific implications for freedom of thought. It is intended to serve as a foundation for further discussion and to inform the initiatives and actions of ODIHR, participating States and other stakeholders.

4 Declaration on the Use of Scientific and Technological Progress in the Interests of Peace and for the Benefit of Mankind, United Nations General Assembly Resolution 3384 (XXX), adopted 10 November 1975.

Introduction

As governments navigate the fast-evolving terrain of artificial intelligence (AI), including its integration in fields such as neurotechnology, a core but often overlooked right demands urgent attention: freedom of thought. Enshrined in the Universal Declaration of Human Rights (UDHR), alongside the right to freedom of conscience and religion, freedom of thought safeguards an individual's ability to think independently, form their own beliefs, make autonomous decisions and contribute new perspectives. This right fosters creativity, innovation and resilient, pluralistic societies.[5]

Freedom of thought also covers the freedom not to reveal one's thoughts, as well as freedom from coercion, from punishment for one's thoughts and from the impermissible alteration of thoughts.[6] Recognizing the fundamental nature of the right, OSCE participating States have committed to protecting freedom of thought, beginning with the Helsinki Final Act.[7]

---

5 International Covenant on Civil and Political Rights, United Nations, General Assembly resolution 2200A (XXI), adopted 16 December 1966, Art. 18:

> "1. Everyone shall have the right to freedom of thought, conscience and religion. This right shall include freedom to have or to adopt a religion or belief of his choice, and freedom, either individually or in community with others and in public or private, to manifest his religion or belief in worship, observance, practice and teaching.
>
> 2. No one shall be subject to coercion which would impair his freedom to have or to adopt a religion or belief of his choice.
>
> 3. Freedom to manifest one's religion or beliefs may be subject only to such limitations as are prescribed by law and are necessary to protect public safety, order, health, or morals or the fundamental rights and freedoms of others.
>
> 4. The States Parties to the present Covenant undertake to have respect for the liberty of parents and, when applicable, legal guardians to ensure the religious and moral education of their children in conformity with their own convictions."

6 This refers to any modification of thoughts done in a manner that runs counter to human rights standards, including through brain alteration and manipulation. See Ahmed Shaheed, Report of the Special Rapporteur.

7 The Helsinki Final Act of the 1st Summit of Heads of State or Government, Conference for Security and Co-operation in Europe (CSCE), 1 August 1975, acknowledges as one of its ten guiding principles the "respect for human rights and fundamental freedoms, including the freedom of thought, conscience, religion or belief". The right was re-affirmed in multiple subsequent OSCE documents (e.g.: Concluding Document of the Third Follow-up Meeting, Vienna, 4 November 1986 to 19 January 1989; Document of the Copenhagen Meeting of the Conference on the Human Dimension of the CSCE, 29 June 1990; OSCE Ministerial Council, Decision No. 4/03, "Tolerance and Non-discrimination", Maastricht, 2 December 2003; OSCE Ministerial Council, Decision No. 3/13 "Freedom of Thought, Conscience, Religion or Belief", Kyiv, 9 December 2013).

Yet, given the rapid development of new technologies, including AI and neurotechnologies, realization of the right to freedom of thought is coming under pressure. As these technologies increasingly shape how we think, communicate and interact, they raise complex legal, ethical and societal questions that require attention and action from participating States.

1. On freedom of thought

Article 18 of the Universal Declaration of Human Rights (UDHR) protects the right to "freedom of thought, conscience and religion".[8] The emphasis is usually placed on religion, with Article 18 often contracted to "freedom of religion or belief" or "religious freedom". Until recently, the inherent protection of freedom of thought was considered so self-evident that a definition seemed unnecessary, despite it being foundational to human dignity and autonomy and to all other human rights. While René Cassin, who co-drafted the UDHR, described freedom of thought as "the origin of all other rights", neither he nor the other drafters elaborated on it further, and little jurisprudence on freedom of thought has followed.[9]

At the same time, in the aftermath of the Second World War, drafters of the UDHR were keenly aware of how this right could be threatened. Attempts at influencing or altering thought, or punishing real or presumed thought and opinion have a long history.[10] People who diverge from the mainstream by thinking differently and challenging existing norms have become targets of coercive actions, which tend to increase with autocracy.[11]

Despite the lack of a definition in international law over what precisely constitutes 'thought',[12] it is, nonetheless, possible to identify its contours, and the attributes necessary to work for its protection in practice.

As part of Article 18 of the UDHR and the International Covenant on Civil and Political Rights (ICCPR), this right has generally been understood in

8 Universal Declaration of Human Rights, UNGA resolution 217 A, 10 December 1948.

9 Ahmed Shaheed, Report of the Special Rapporteur.

10 From Nazi propaganda that led to the Holocaust to totalitarianism and authoritarianism and other situations where states have taken measures that interfere with freedom of thought, such as interventions to prevent violent extremism and radicalization leading to terrorism. See, for example, 'This is the Thought Police': The Prevent Duty and Its Chilling Effect on Human Rights, (London: Amnesty International, 2023); UN Special Rapporteur (UNSR) on Freedom of Religion or Belief (FoRB) Ahmed Shaheed, Report on the relationship between freedom of religion or belief and national security, A/73/362, 5 September 2018, para. 43.

11 71% of the world population was living under autocratic regimes in 2023 compared to 50% in 2003; in 2024, 42 countries (hosting 35% of the world population) were going through 'autocratization' and 18 countries (hosting 5% of the world population) were going through democratization. See Marina Nord, Martin Lundstedt, and Staffan I Lindberg, Media Freedom, Democracy, and Security, Research Report, OSCE/RFoM, 15 July 2024, p. 6.

12 Ahmed Shaheed, Report of the Special Rapporteur, para. 11.

relation to freedom of conscience and freedom of religion or belief. However, freedom of thought does not only apply to religious thought. The UN Human Rights Committee, which provided an authoritative interpretation of ICCPR Article 18 in General Comment 22, explained that it "encompasses freedom of thought on all matters".[13]

Importantly, no derogation can be made from Article 18, meaning that states cannot suspend their obligations in relation to this right, even under conditions of "public emergency which threatens the life of the nation".[14] Freedom of thought is generally understood to be an absolute and unconditional right. General Comment 22 notes that Article 18 "does not permit any limitations whatsoever on the freedom of thought and conscience or on the freedom to have or adopt a religion or belief of one's choice".[15] These two freedoms, together with the right to hold opinions without interference (ICCPR Article 19.1), form what is known in human rights law as the *forum internum* — or the 'inner sanctum' of the human mind.[16]

The *forum internum* is held in contrast to the *forum externum*, which concerns the expression of thoughts, convictions or beliefs. *Forum externum* rights are not absolute.[17] Rather, they may be subject to limitations by the state "as are prescribed by law and are necessary to protect public safety, order, health, or morals or the fundamental rights and freedoms of others" (and in the case of Article 19.3, on the additional grounds of protecting national security).

This formulation distinguishes between the aspects of Articles 18 and 19 that have absolute protection and those which may be restricted under

13 General Comment 22, "Article 18 (Freedom of Thought, Conscience or Religion)", UN Human Rights Committee, Forty-eighth session, 27 September 1993, para. 1.

14 A number of other rights cannot be derogated from, including Art. 6 – the right to life; Art. 7 – the right not to be subjected to torture or to cruel, inhuman or degrading treatment or punishment; Art. 8 (paras. 1 and 2); the right not to be held in slavery or servitude and the obligation to prohibit slavery and slave trade in all their forms.

15 Ahmed Shaheed, Report of the Special Rapporteur, para. 3.

16 *Ibid*., para. 2.

17 ICCPR, Art. 18.3: "Freedom to manifest one's religion or beliefs may be subject only to such limitations as are prescribed by law and are necessary to protect public safety, order, health, or morals or the fundamental rights and freedoms of others." ICCPR, Art. 19.3: "The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:

(a) For respect of the rights or reputations of others;

(b) For the protection of national security or of public order (ordre public), or of public health or morals."

certain circumstances. However, to regard freedom of thought purely in terms of the *forum internum* risks giving it a passive character. Thought is not merely a driver of human activity, but an activity itself that should be allowed to proceed unhindered.

As with all human rights, freedom of thought often overlaps with other rights and freedoms with significant interrelationships and co-dependencies. Perhaps one of the most fundamental examples is its relationship with privacy, a right which can be limited in certain situations, in terms of protecting (mental) space to generate thought. This overlap also illustrates the need to further outline the scope of freedom of thought and to better clarify its protections, especially given the absolute nature of the right. For example, under what circumstances do mental privacy and integrity warrant absolute protection under the right to freedom of thought? And when would qualified protection under the right to privacy suffice?[18]

Freedom of thought also needs to be understood in relation to freedom of conscience and freedom of religion or belief. Conscience, which, as the 'inner tribunal',[19] allows us at its most basic level to freely decide what is right and wrong, only functions if it flows from autonomous thought, in the sense that each person's autonomy to shape thoughts is not unduly interfered with. Similarly, there is a risk that adherence to any religion or belief system could be coercive in nature if there is no freedom of thought.

The process of generating thought does not occur in a vacuum, but in interaction with other people, spaces and information, where communities

18 Academic literature has advanced a number of criteria to help make the distinction between situations that warrant absolute and qualified protection, in relation to freedom of thought and privacy respectively. One criterion takes into account the mental effects of an interference (chiefly in terms of 'severity' and 'significance') alongside other relevant factors, such as the importance of personal interest in not revealing certain types of mental content (e.g., sexual orientation, political opinion, religious beliefs), or the method of interference by which a person's mental state is revealed or changed (e.g., indoctrination or brainwashing). The level of control over one's own thoughts after the interference is also relevant here, as well as whether the method undermines such control or bypasses it. Another criterion looks at the characteristics of the victim, where vulnerability is important; children, the elderly, people with mental disabilities or people deprived of liberty or in other custodial settings are particularly vulnerable. Power differentials are important in this sense and could also include situations in the workplace. Another criterion is the context in which the interference took place, which warrants a case-by-case assessment. See Sjors Ligthart and Naomi van de Pol, Freedom of Thought: Absolute Protection of Mental Privacy and Mental integrity? Considering the Case of Neurotechnology in Criminal Justice, in Patrick O'Callaghan and Bethany Shiner (eds.), *The Cambridge Handbook of the Right to Freedom of Thought*, (Cambridge: Cambridge University Press, 2025).

19 Sofie Møller, *Kant's Tribunal of Reason, Legal Metaphor and Normativity in the Critique of Pure Reason,* (Cambridge: Cambridge University Press, 2020), Chapter 6 - Moral Conscience as the Practical Inner Tribunal, pp. 96-112.

also play a role. From this perspective, there is an important overlap with the rights contained in Article 19 — freedom of opinion and expression.[20] Thought and expression exist in a "perpetual feedback loop".[21] They also remain distinct and are protected differently in the international human rights framework. Thought refers to the activity of thinking, from stimulus to output, whereas opinion can be understood as one of the provisional outputs of thought.[22]

> "(…) humans develop and flourish in their interactions with other human beings and a nurturing material and cultural environment, (…) autonomy is not just individual, but also relational as it arises from and impacts one's interactions and belonging with the community."
>
> — Draft text of the Recommendation on the Ethics of Neurotechnology, UNESCO, 9 April 2025, paras. 16 and 44.

In a seminal 2021 report (and the first substantive commentary within the UN system on the right to freedom of thought), UN Special Rapporteur on Freedom of Religion or Belief (UNSR), Ahmed Shaheed, outlines four possible attributes of the right:

- Freedom not to reveal one's thoughts
- Freedom from punishment for one's thoughts, real or inferred
- Protection from impermissible alteration of thought
- States' positive obligation to create an enabling environment for freedom of thought

20 Other relevant rights include: The right of everyone to the enjoyment of the highest attainable standard of physical and mental health (Art. 12 of the International Covenant on Economic, Social and Cultural Rights — ICESCR); the right of everyone to education — also in the sense that "education shall be directed to the full development of the human personality and the sense of its dignity" (Art. 13, ICESCR); the right to take part in cultural life, and enjoy the benefits of scientific progress and its applications, the right to benefit from the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author, or states undertaking to respect the freedom indispensable for scientific research and creative activity (Art. 15, ICESCR), or the prohibition to compel anyone to belong to an association (Art. 20.2 of the UDHR).

21 Ahmed Shaheed, Report of the Special Rapporteur, para.18.

22 *Ibid*., para. 21; See also Christoph Bublitz, 'The Mind and Conscience are the Person's Most Sacred Possessions': The Origins of Freedom of Thought in the Universal Declaration of Human Rights and the International Covenant on Civil and political Rights, in Patrick O'Callaghan and Bethany Shiner (eds.), *The Cambridge Handbook of the Right to Freedom of Thought*, (Cambridge: Cambridge University Press, 2025).

While the first two are more straightforward, the protection from impermissible alteration of thought generally refers to the prohibition of interference with mental autonomy. This assumes the absence of coercion, thought modification through brain alteration and manipulation.

Freedom of thought does not, however, guarantee 'immunity' from other people's thoughts, or from everyday processes of persuasion. The UNSR further proposes, among other factors, possible criteria to identify undue forms of manipulation, which should be assessed on a case-by-case basis:

- Consent (from a person able to give it, free and informed)
- Concealment and obfuscation (if a 'reasonable person' would have been aware of the intended influence)
- Asymmetrical power between the influencer and rights holder and how this was used
- Harm in intent or effect

The fourth attribute, the obligation to create an enabling environment for freedom of thought, is linked to other rights, including the rights to access information, communication and education.[23]

> "We need freedom of thought to combat climate change, racism and global poverty, and to fall in love, laugh and dream. The right to freedom of thought is an individual right, but it is crucial to the cultural, scientific, political and emotional life of our societies. Freedom of thought gives us the chance to think ugly thoughts and push them away before we act on them or let them take root; it allows us to choose how we behave to others, to moderate our speech according to the context and the audience and to be ourselves. Freedom of thought lets us imagine new futures without having to prove them first, it keeps us dynamic and adventurous; it keeps us safe; and above all, it keeps us human."
>
> — Susie Alegre, *Freedom to Think: Protecting a Fundamental Human Right in the Digital Age*, (London: Atlantic Books, 2022), Introduction.

23 *Ibid.,* paras. 25-47.

2. On artificial intelligence

There is no universally accepted definition of 'AI' in international law. The European Union (EU) AI Act,[24] defines an "AI system" as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments".[25] A similar definition is included in the Council of Europe's Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.[26]

AI relies on algorithms[27] written in computer code that process datasets by collecting and transforming raw data to produce outputs.[28] Unlike an algorithm, where the process of decision-making is predefined, AI 'makes decisions' on the basis of data analysis.[29] It identifies patterns in the data, 'learns' from them and makes predictions, which it can then adapt as it continues to collect data.[30]

The potential applications of AI are nearly endless, and the consequences for people's lives vary significantly. AI operates on datasets that capture social realities and personal data, including in health care, employment, banking, shopping or law enforcement settings. To illustrate, in terms of health care, it can analyse sets of medical data to predict health outcomes that inform medical decisions. AI is also used to analyse large online datasets on shopping behaviour to infer individual preferences. It can then correlate this information

24 European Union Artificial Intelligence Act, Regulation (EU) 2024/1689, Official Journal version of 13 June 2024.

25 *Ibid.* Art 3 (1).

26 Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Council of Europe, adopted on 5 September 2024.

27 Algorithms are a set of instructions aimed at solving a specific problem. The processes through which they operate are complex and varied. See, for example, Andrew Williams, What Is an Algorithm? Defining And Applying Algorithms, *Forbes*, 12 January 2024.

28 UN Special Rapporteur (UNSR) on the promotion and protection of the right to freedom of opinion and expression (FoOE), David Kaye, A/73/348, 29 August 2018.

29 Kaya Ismail, AI vs. Algorithms: What's the Difference?, *CMSWire,* 26 October 2016.

30 Office of Communications, College of Education, Traditional AI vs. Generative AI: What's the Difference?, University of Illinois Urbana-Campaign, 11 November 2024.

with personal and contextual characteristics (e.g., age, geographic location, gender, past purchases and shopping behaviour) to create detailed customer profiles. Companies use these profiles for market segmentation, enabling them to target individuals with tailored products or services designed to increase sales or engagement with specific content. Engagement leads to new user data being harvested.

Although users are more or less aware of the use of their data or the influence on their choices, these processes still raise questions over freedom of thought and the potential for manipulation. Equally, users do not have sufficient information to participate in debates over how new technologies can be used in the public interest.

Generative AI

Generative AI (GenAI) is a new generation of AI development. It refers to artificial intelligence that 'creates' content, which is mostly used in the context of responding to prompts provided through natural-language interfaces.[31] 'Traditional' AI produces outputs on the basis of existing content. GenAI generates 'new' content, on the basis of existing content.[32] This can include text, images, videos, music and even software code. GenAI is trained on immense amounts of data. To generate its outputs, it conducts sophisticated analysis of patterns and relationships in the data, such as word sequences or pixel arrangements.[33] It is based on, and has been made possible by what is known as large language models and deep learning;[34] these had been developed and improved over a number of years before ChatGPT became available for everyone to experiment and work with.

Once OpenAI made ChatGPT — its GenAI model trained on internet data — publicly available, it became the fastest growing app, sparking a race among major tech companies to develop GenAI tools and integrate them into

31 Fengchun Miao, Wayne Holmes, Guidance for generative AI in education and research, (Paris: UNESCO, 2023).

32 Office of Communications, College of Education, Traditional AI vs. Generative AI: What's the Difference?.

33 Miao and Holmes, Guidance for generative AI in education and research, p. 8.

34 Cambridge English Dictionary Online, Entry for Large Language Model.

various products and platforms.[35] Notably, due to its fast-paced roll-out, there has been a sharp rise in AI-generated content, which, by adding to human-generated content, is changing the overall nature of the data on the Internet and, therefore, the available knowledge.[36]

One area of concern, including to the industry, is that, although this may change with new technological advancements, research has shown that GenAI models collapse when repeatedly trained on artificially-generated data, as opposed to human-generated data.[37]

> "Model collapse is a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality."
>
> — Ilia Shumailov, Zakhar Shumaylov, et.al., AI models collapse when trained on recursively generated data, *Nature*, 631, 24 July 2024, pp. 755-759.

In early 2025, the launch of DeepSeek, a Chinese-developed GenAI model, offered high technical performance at much lower cost, including in environmental terms. DeepSeek was launched as open source, enabling anyone to download, copy and build upon it without having to go through the significant costs of building a model from scratch,[38] thus making new scientific discovery publicly available.[39] However, its emergence has also raised rights concerns related to privacy and censorship,[40] and some EU countries have banned its use on grounds of data protection.[41] At the time of writing, the

35 Other examples of GenAI models include: Claude (Anthropic), Llama (Meta), Gemini (Google), Le Chat (Mistral) and others. Major developments come from big companies, but the applications are not restricted to Big Tech.

36 See Jason Koebler, Project Analyzing Human Language Usage Shuts Down Because 'Generative AI Has Polluted the Data', *404media,* 19 September 2024.

37 See Ilia Shumailov, Zakhar Shumaylov, et.al., AI models collapse when trained on recursively generated data, *Nature*, 631, pp. 755-759, 24 July 2024.

38 Charlotte Edmond, What is open-source AI and how could DeepSeek change the industry, World Economic Forum*,* 5 February 2025.

39 Alex He, DeepSeek and China's AI Innovation in US-China Tech Competition, Centre for International Governance Innovation*,* 11 April 2025.

40 Robert Booth and Dan Milmo, Experts urge caution over use of Chinese AI DeepSeek*, The Guardian,* 28 January 2025.

41 Hakan Ersen and Miranda Murray, DeepSeek faces ban from Apple, Google app stores in Germany, *Reuters,* 27 June 2025.

impact of DeepSeek on competition in the field and the spread of GenAI models is yet to be fully seen.[42]

Whether truly close or not,[43] large tech companies are also talking about concrete steps towards the next stage of artificial intelligence: Artificial General Intelligence (AGI). AGI is expected to operate autonomously and exhibit human-like agency.[44] This has implications for freedom of thought, as such systems could influence, enforce or undermine this right. It is, therefore, essential that governments, developers and regulators ensure that potential AGI is designed and governed in accordance with human rights standards.

Corporate influence in the age of AI

The economic and regulatory environment in which new technologies have been operating during past decades has already had a profound impact on democracy and human rights. The use of AI by large technology companies, often called 'Big Tech', that operate major internet platforms (including search engines and social media) has significantly transformed the space for public discourse. This transformation has changed democratic debate and the way people form their opinions, make choices and reach decisions, with direct implications for freedom of thought.[45] As corporations and governments gain ever-greater access to large amounts of data about individuals' behaviours and preferences, the potential to influence or even manipulate people's innermost thoughts has also increased. Regardless of whether they are funded publicly or privately, AI-based technologies, with a few exceptions, currently

42 Hamilton Mann, Seek Deeper On DeepSeek For Artificial Integrity Over Intelligence, *Forbes,* 28 January 2025; Robert Booth and Dan Milmo, Chinese AI chatbot DeepSeek censors itself in realtime, users report, *The Guardian*, 28 January 2025; Cade Metz, How Did DeepSeek Build Its A.I. With Less Money?, *The New York Times*, 12 February 2025.

43 Kate Brennan, Amba Kak and Dr. Sarah Myers West, Artificial Power: 2025 Landscape Report, AI Now Institute, 3 June 2025.

44 Tom Allen, Real-world agentic AI is 'complex' cautions Google's Demis Hassabis, Computing, 17 March 2025; Will Knight, Google's AI Boss Says Gemini's New Abilities Point the Way to AGI, *Wired,* 28 May 2025; Robert Booth, Meta to announce $15bn investment in bid to achieve computerised 'superintelligence', *The Guardian,* 11 June 2025.

45 See a detailing of the argument in Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, (New York, Public Affairs, 2019). See also the work of the OSCE RFoM on AI and Freedom of Expression.

lack effective, human rights-based public regulatory frameworks to ensure prior impact assessment or subsequent accountability.

Neurotechnologies are also developing rapidly and becoming increasingly sophisticated through their convergence with AI. These technologies open up the possibility to infer and influence people's thoughts, since such technologies operate in relation to the brain and the neural circuit. The commercialization of devices that harvest neural data increases the possibilities for profiling and surveillance. This is directly relevant to freedom of thought. While such devices offer potentially ground-breaking solutions, especially in the field of medicine, they may allow and be presumed to offer unprecedented insights into what was long considered an inviolable inner sphere — the sanctity of the mind. When used for profit, these technologies have the potential to feed into and enhance the power and sophistication of the influence over individual choices that large online platforms exert, which poses new ethical and human rights challenges.[46]

AI technologies are primarily developed and deployed by private sector actors, who operate within commercial frameworks where generating profit is essential for sustainability. As a result, these technologies are largely shaped outside democratic oversight and the traditional 'social contract' between governments and citizens, despite their significant impact on issues of public interest. This disconnect means that the governance structures currently in place are not currently able to address adequately the broad influence that these technologies have on society. It also means that the commercial priorities of these private actors may or may not align with the wider objectives of public welfare, creating challenges in ensuring that AI development always serves the common good.

Given AI's profound impact on democratic institutions, human rights and the ways people interact, it is important to address the question of who should develop and deploy such technologies, under what conditions and regulation, and to what ends. Beyond the dominant Big Tech companies, there are myriad start-ups and other competitors in the AI field.[47] Industry standards, and especially ethical standards, are still evolving, and actual

46 See, for example, Nita A. Farahany, *The Battle for Your Brain. Defending the Right to Think Freely in the Age of Neurotechnology*, (New York: St Martin's Griffin, 2023).

47 Rashi Shrivastara (ed.), Forbes 2025 AI 50 List - Top Artificial Intelligence Companies Ranked, *Forbes,* 10 April 2025; Kate Brennan et al, Artificial Power: 2025 Landscape Report.

practices often depend on the economic power or business culture of individual companies or the environment in which they operate.

States must take responsibility for ensuring that the 'AI revolution' safeguards everyone's human rights and fundamental freedoms. In so doing, it is important for states to promote a level playing field that encourages fair competition among all actors and prevents a race to the bottom. States have an ongoing duty to create and protect an enabling environment for freedom of thought and all other human rights, while carefully monitoring Big Tech's influence to ensure accountability and alignment with public interest.[48]

48 Other pertinent areas with specific implications for freedom of thought, not approached here, include the use of AI for distinct surveillance purposes, especially for national security or defence purposes, at borders and in migration control, during detention, by the military or as part of authoritarian regimes. See also Digital technologies at borders: A threat to people on the move, UN OHCHR, 9 October 2023.

3. The human rights implications of thinking within the algorithm

AI, bias and thought

AI outputs depend on the purpose and design of the AI system, the quality of the data and how the AI is deployed. A specific AI system will reflect both the biases and blind spots of those involved in its creation, as well as those embedded in the datasets used to build and train it. Furthermore, it may be deployed for different purposes than it was created and tested for, while a number of inherent issues may appear in the machine learning algorithms as such.[49] In fields traditionally involving social interaction (e.g., banking, social benefits, housing, employment), if the datasets already contain biases, for example, this will create particularly problematic consequences, as AI systems have power at scale to proliferate and potentially exacerbate such biases.[50] AI systems also commonly reflect the values and realities of the global north, making AI more likely to misrepresent other regions. All of these issues carry risks to human rights, including the right to freedom of thought. However, efforts must continue to address this issue and reduce biases by, among others, promoting diversity throughout the AI lifecycle, working to fix technical challenges and, in some cases, not using certain types of AI if the risk is too high, if thought is to be enhanced, diversity protected and access to varied perspectives ensured.

> "A feedback loop occurs when predictions made by a system influence the data that are used to update the same system. It means that algorithms influence algorithms, because their recommendations

49 Bias in Algorithms: Artificial Intelligence and Discrimination, EU Agency for Fundamental Rights (FRA), (Vienna, FRA, 2022).

50 For instance, predictive policing algorithms can be influenced by a number of factors, including if the dataset is based on prior recorded crime rates. These rates can in turn be influenced by prejudice against certain minorities who may live in certain areas, which then show up more often in the datasets as areas where crime rates are high than they would in the real (recorded and non-recorded) crime rates. The algorithmic predictions can thus become biased against the specific minorities, whose neighbourhoods may end up being overpoliced. For a comprehensive explanation on bias in AI predictive policing systems, see FRA, Bias in Algorithms, p. 31. The EU AI Act prohibits individual predictive policing AI systems solely based on profiling or personality traits, except when used to augment human assessments based on objective, verifiable facts directly linked to criminal activity (see also Chapter 6 below).

and predictions influence the reality on the ground. This reality then becomes the basis for data collection to update algorithms. Therefore, the output of the system becomes the future input into the very same system. Any bias in algorithms can therefore potentially be reinforced over time and exacerbated."

— EU Agency for Fundamental Rights, Bias in Algorithms: Artificial Intelligence and Discrimination, (Vienna, FRA, 2022), p. 8.

When human interactions and decisions are replaced by AI systems, or there is an over-reliance on such systems, the environment for informed decision-making is distorted by pre-existing or system-generated inequities and inaccuracies, which are then amplified by the system. Where biased social media content or AI profiling informs political and other social activity, people with specific protected characteristics or behaviours are often most immediately affected.[51] Ultimately, however, this affects society as a whole.

Given their complexity and size, curating datasets to avoid bias, rather than trying to reduce it, is a challenging task. Even when the most well-intentioned efforts are made to remove bias from AI tools, errors in the process can, by themselves, lead to (further) human rights violations. For example, social media platforms using AI for moderation — to filter, flag, remove or limit the spread of illegal and offensive content — are already operating on a vast scale. As it tries to navigate the complexities of human communication on these platforms, AI often fails to understand either context or nuance (e.g., satire), which limits how well they moderate human communication in general, let alone the complexities of bias.[52] Additionally, meaning varies according to circumstances, shared understanding and language, and evolves over time. When trying to use AI to moderate for bias, social media platforms may inadvertently censor marginalized communities for engaging with and using certain language, including protected counter-speech that opposes or challenges harmful and hateful speech.[53] At the same time, content that uses 'coded' language can avoid detection. These challenges are more pronounced for content in languages other than

51 Mehnaz Rafi, When AI plays favourites: How algorithmic bias shapes the hiring process, *The Conversation*, 14 October 2024; see also ICO considers uses of neurotechnology in employment in the UK, Withers, 31 July 2023.

52 Spotlight on AI and Freedom of Expression – A Policy Manual, OSCE/RFoM, 20 January 2022.

53 FRA, Bias in Algorithms: Artificial Intelligence and Discrimination.

English, where companies have invested fewer resources in both moderation and fact-checking.[54] Efforts to curate data for bias, which mainly rely on AI, should always include the sufficient engagement of diverse and well-informed humans from the outset of the AI life cycle.

It is worrying that, currently, social media companies are reversing their policies on content moderation and protections related to trust and safety, including hate mitigation, harassment and false content. They have significantly reduced the number of staff required for proper moderation[55] and are therefore getting worse, not better, at countering harmful content.[56] Given the damage such material does to marginalized communities, (including the real-life danger that online hate can provoke), efforts to mitigate bias should, if anything, be increased.[57]

One particularly important characteristic of the digital environment is that English dominates, not reflecting the diversity of languages spoken across the world. This has exacerbated the digital divide in terms of access to the internet for those who do not speak English or other common internet languages.[58] GenAI adds to this issue, as it is trained in a limited number of languages and therefore fills the information environment primarily in those languages. This trend threatens language diversity and is unlikely to change without significant investment by companies.[59]

Native languages are not just essential to cultures and identities, but represent a constitutive part of how people organize their thoughts in the *forum internum* and how people's worldviews are shaped. On the positive side,

54 Content Moderation in a New Era for AI and Automation, Meta Oversight Board, September 2024; see also OSCE/RFoM, Spotlight on AI and Freedom of Expression – A Policy Manual.

55 David Evan Harris and Aaron Shul, Generative AI, Democracy and Human Rights, Centre for International Governance and Policy, Policy Brief No. 12, February 2025.

56 Joel Kaplan, More Speech and Fewer Mistakes, Meta, 7 January 2025; Clare Duffy, Calling women 'household objects' now permitted on Facebook after Meta updates its guidelines, *cnn.com*, 8 January 2025; Dia Kayyali, Meta's Content Moderation Changes are Going to Have a Real World Impact. It's Not Going to be Good, *TechPolicy.press*, 9 January 2025, Adrian Kopps, Four key policy changes of X under Musk, Digital Society Blog, Alexander von Humboldt Institut für Internet und Gesellschaft, 28 October 2024.

57 Hate Speech, United Nations-dedicated webpage; X's design and policy choices created fertile ground for inflammatory, racist narratives targeting Muslims and migrants following Southport attack, Amnesty International, 6 August 2025.

58 Internet Access: UNESCO and ICANN join forces to improve linguistic diversity online, UNESCO, 27 February 2025.

59 Viorica Marian, AI could extinguish languages and ways of thinking, *The Washington Post*, 19 April 2023.

AI has been used successfully, among other public benefit purposes,[60] to document, preserve and save endangered languages, and even revive lost ones.[61] Central to this is the human rights ethos and understanding — or lack thereof — shaping the development, deployment and use of AI.

Without appropriate care to ensure diversity and representation, including by closing the digital divide, an AI future where a myriad of power imbalances is (re-)created through the use of new technologies operating at scale risks making worldviews not represented by AI disappear.

Human oversight, transparency and accountability

From the perspective of ensuring human rights compliance, the level of human intervention in decision-making assumes critical importance. Fully automated decision-making means that there is no human intervention (including human intuition or judgement) involved in how AI delivers an output. Furthermore, the complexity of GenAI makes it difficult, if not impossible, to understand or explain how a given output was reached. There is an additional complication with human-machine interaction, namely 'automation bias': studies have shown that humans generally trust machines more than they trust people and, therefore, more easily accept AI-generated output.[62] Freedom of thought and, indeed, meaningful human agency are undermined when it becomes more difficult to verify and question output, especially in a consequential situation, even if humans are involved in deciding how to apply or act upon those outputs.[63] The level of risk varies depending on the field where AI is employed. For example, doctors should fully understand the AI they use and be trained to overcome the automation bias, systematically evaluating the AI output for each patient since the risk, in terms of negative health outcomes, is very high.[64]

60 See the UN's AI for Good platform and Stéphanie Bascou, « IA d'intérêt public » : cette ONG veut créer un label éthique, similaire au « Commerce équitable », et facilement identifiable pour les consommateurs, (AI in the public interest: this NGO wants to create an ethical label, similar to "Fair Trade", and easily identifiable for consumers), *01.net*, 5 October 2024, (in French).

61 AI: The Unexpected Hero in the Battle to Save Dying Languages, apolitical website, 20 October 2024.

62 Chris Baraniuk, Why we place too much trust in machines, *BBC,* 20 October 2021.

63 For a comprehensive explanation on predictability and understandability, see Arthur Holland Michel, The Black Box, Unlocked, UNIDIR, 22 September 2020.

64 The application of artificial intelligence in healthcare and its impact on the "patient-doctor" relationship, Council of Europe, Steering Committee for Human Rights in the Fields of Biomedicine and Health (CDBIO), September 2024, p. 20.

Frequently, the ability to verify and challenge output is further challenged by claims of intellectual property rights on the code, parameters and/or other inner workings of AI. These arguments are used to avoid disclosing how a certain result was achieved. When transparency and the ability to decide over the information environment (e.g., what social media algorithms display in feeds) are removed, it further complicates informed decision-making or even building a proper understanding of one's environment.

> "Understandability refers to the degree to which any given system can be understood by any given person. Whereas a system's predictability relates to the question *What will the system do?* understandability relates to the question *Why does it do it?*"
>
> — Arthur Holland Michel, The Black Box, Unlocked, UNIDIR, 22 September 2020, p. 9.

The inability to understand how AI reaches decisions that affect human rights — either because the system is too complex or because its workings and the means to control it are not publicly available information — raises serious concerns about accountability and transparency; two crucial principles for the proper functioning of justice systems and democracies.

> "1. Digital technologies are dramatically transforming our world. They offer immense potential benefits for the well-being and advancement of people and societies and for our planet. (…) 3. We recognize that the pace and power of emerging technologies are creating new possibilities but also new risks for humanity, some of which are not yet fully known. We recognize the need to identify and mitigate risks and to ensure human oversight of technology in ways that advance sustainable development and the full enjoyment of human rights."
>
> — UN Global Digital Compact, annexed to the United Nations General Assembly resolution A/RES/79/1 on the Pact for the Future, adopted 22 September 2024.

The way to mitigate these risks is to build AI differently,[65] in a responsible and human rights-centred manner, ensuring regulatory oversight and adequate understandability, transparency and accountability. Furthermore, meaningful human oversight is crucial to ensuring that AI systems help deliver the best results for people and do not undermine their rights, including freedom of thought.

65 Recommendation on the Ethics of Artificial Intelligence, UNESCO, 23 November 2024; see also Rumman Chowdrhury and Dhanya Lakshmi, "Your opinion doesn't matter, anyway": exposing technology-facilitated gender-based violence in an era of generative AI, (Paris: UNESCO 2023); or the Resources page of the organization Humane Intelligence.

4. Freedom of thought in reconfigured civic space

The new civic space architecture impacting freedom of thought

Genuine freedom of opinion and expression depends on genuine freedom of thought.[66] As explained above, both thought and opinion are part of the *forum internum* — the private inner domain of each individual — an area that enjoys absolute protection from interference under international human rights law. Conceptually, in democratic theory, free thought takes shape and evolves in conversation, as individuals freely exchange opinions and share information.[67] In this ideal civic space, it must be possible to express, challenge and test all ideas and opinions, to ensure that the best ideas and truths can ultimately prevail. The ability to have this dynamic exchange, where ideas can meet and confront each other, is why freedom of expression (i.e., the manifestation of thought and opinion) enjoys strong protection and support; it is critical for the healthy public debate that is essential to democracy. Public interest journalism plays a critical role in enabling informed democratic debate, as well as the possibility, among others, to freely associate and organize. However, there are also limits. For example, states should restrict 'hate speech' that reaches a certain threshold of severity, in line with international human rights law.[68]

With the rise of Big Tech, much communication has shifted onto social media platforms, which now mediate human interactions and act as gatekeepers to news/information. These platforms use powerful machine-learning algorithms to reorganize content in ways that have implications for freedom

66 Sjors Ligthard, Christoph Bublitz, Thomas Douglas, Lisa Forsberg and Gerben Meynem, Rethinking the Right to Freedom of Thought: A Multidisciplinary Analysis, in *Human Rights Law Review,* 2022, 22, pp. 1-14.

67 See also Simon McCarthy-Jones, *Freethinking: Protecting Freedom of Thought Amidst the New Battle for the Mind*, (London: Oneworld, 2023).

68 ICCPR, Article 20:

"1. Any propaganda for war shall be prohibited by law.

2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

of thought, as they fundamentally change at scale how information, opinions and ideas circulate, as compared to an information environment largely organized by traditional (public interest) media.[69]

Online, algorithm-based recommender systems are primarily shaped by commercial considerations. These systems are optimized to maximize user engagement or drive purchases rather than to prioritize accuracy, diversity or public interest goals, including media pluralism and public interest journalism.[70] Maximizing engagement and driving purchases involves using past behaviour and individual profiling to predict what content users are most likely to engage with. As a result, users have limited control over what they see, even if they still retain autonomy over whether to engage or not. They may also be unaware that the material they see differs from that seen by others, including those in their close circle, due to algorithmic personalization.

This model places significant influence over each person's information ecosystem in the hands of a few platforms. It has contributed to the fragmentation of the news media landscape, weakening independent journalism and potentially eroding a shared sense of reality, which has long shaped individuals' interactions, choices and democratic engagement.[71]

---

69 "Unlike traditional algorithms, which are hard-coded by engineers, machine-learning algorithms 'train' on input data to learn the correlations within it. The trained algorithm, known as a machine-learning model, can then automate future decisions. An algorithm trained on ad click data, for example, might learn that women click on ads for yoga leggings more often than men. The resultant model will then serve more of those ads to women." See Karen Hao, The Facebook whistleblower says its algorithms are dangerous. Here's why, *MIT Technology Review*, 5 October 2021.

70 OSCE/RFoM, Spotlight on AI and Freedom of Expression – A Policy Manual.

71 OSCE/RFoM, Spotlight on AI and Freedom of Expression – A Policy Manual; and, more broadly, the work of the OSCE RFoM on Media and Big Tech; see also Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes, Council of Europe Committee of Ministers, adopted on 13 February 2019: "8. Contemporary machine learning tools have the growing capacity not only to predict choices but also to influence emotions and thoughts and alter an anticipated course of action, sometimes subliminally. (…) 9. Fine grained, sub-conscious and personalised levels of algorithmic persuasion may have significant effects on the cognitive autonomy of individuals and their right to form opinions and take independent decisions. (…) [Council of Europe's] central pillars of human rights, democracy and the rule of law are grounded on the fundamental belief in the equality and dignity of all humans as independent moral agents. (…) the Committee of Ministers: - draws attention to the growing threat to the right of human beings to form opinions and take decisions independently of automated systems, which emanates from advanced digital technologies. Attention should be paid particularly to their capacity to use personal and non-personal data to sort and micro-target people, to identify individual vulnerabilities and exploit accurate predictive knowledge, and to reconfigure social environments in order to meet specific goals and vested interests;"

"The methods by which online platforms curate content through recommender systems are not transparent, and they are very rarely subject to public and/or state scrutiny. (…). Content recommender systems may also have unintended consequences from the perspective of broader societal objectives and can negatively shape and interfere with the absolute right to freedom[s] of thought and opinion. In addition, the processes of internet intermediaries' recommender systems typically exclude individual users' choice, control and agency – prerequisites to ensuring individual autonomy in seeking and imparting a variety of information and ideas. (…) There exists ample evidence that online platforms' opinion power has the ability to steer and amplify certain public narratives and types of discourse over others. For countries with fragile or oppressive political systems, this opinion power, coupled with algorithmic amplification, can have disastrous consequences for individual enjoyment of human rights."

— Spotlight on AI and Freedom of Expression – A Policy Manual, OSCE/RFoM, 20 January 2022, pp. 64- 78.

The OSCE Representative on Freedom of the Media has done extensive work exploring and explaining the impact of AI on the digital marketplace of ideas, public interest media, freedom of opinion and expression and democratic processes, offering comprehensive recommendations to participating States.

See more in OSCE RFoM Spotlight on Artificial intelligence and Freedom of Expression — SAIFE Resource Hub.

As AI capabilities develop, advertising that relies on an AI-driven data-harvesting business model, which analyses individual and collective behaviour and characteristics, is becoming more effective. This system tracks users across sites and devices, delivering targeted content, including news, to maximize engagement and drive purchases.

The advent of social media has undeniably enabled new contact and communication to take place, facilitated popular mobilization on issues of common concern and given a voice to those under censorship. Compared to traditional public spaces, however, social media platforms modify how engagement with one another and groups happens. This communication framework is set up to leverage human psychological biases — especially

around 'in-group' vs. 'out-group' dynamics and the human need to belong — to deliver tailored content according to criteria that are largely not publicly available (i.e., how the algorithms that drive content are set up and for what purposes). Platforms have transformed much of the public space and how we are able to engage with it, while at the same time maintaining a strong illusion of choice.[72] Through subtle, yet systematic effects, people become susceptible to influence and manipulation of thought and behaviour, undermining the idea of thinking for oneself and diminishing our agency.[73]

Arguably, the dominant position of the large technology companies leaves individuals with little choice but to accept the surveillance model and to be on the platform. It has become a requirement for many everyday social and professional interactions, particularly where group activities or other important information is only communicated via one platform. This has consequences for freedom of thought. By submitting to companies' algorithms to filter a significant part of their experience of the world, people risk relinquishing control over their selfhood. The very knowledge of being surveilled can also influence what humans think and how they act, potentially in fear of consequences.[74] Importantly, the EU has started to take steps

72 One, often evoked, effect of the new information environment, especially as related to political polarization, is that of creating 'echo chambers' — the result of people's tendency, followed by the actual possibility of choice on social media, to select to engage with people like them whose ideas they agree with. While the literature is divided on the existence of echo chambers, it is also very limited geographically and linguistically, and is generally produced without appropriate access to the data necessary for proper analysis. A large, collaborative, meta-research study looking at political engagement concluded there is evidence of people immersing themselves in homogenous, partisan networks when on social media. But the picture is much more complex. The authors also highlight that, while social media may offer exposure to more news information and ideas than one would otherwise encounter in their offline world, when online, different opinions and information are perceived from a polarized position; people react to opposing views, not by entering into a democratic debate, but by seeking approval from an online community — an 'in-group' against the 'out-group' — which finally leads to even more entrenched positions and polarization. See Jonathan Haidt and Chris Bail (ongoing), Social Media and Political Dysfunction, A collaborative review, unpublished manuscript, New York University, Chapter. 2.4).

73 The term 'surveillance capitalism' was popularized by Shoshana Zuboff, who describes it as a new economic logic centred on the extraction of personal data for the purpose of shaping user behaviours. " (…) surveillance capitalists discovered that the most-predictive behavioral data come from intervening in the state of play in order to nudge, coax, tune, and herd behaviour toward profitable outcomes. Competitive pressures produced this shift, in which automated machine processes not only know our behavior, but also shape our behaviour at scale. With this reorientation from knowledge to power, it is no longer enough to automate information flows about us, the goal is to automate us." Shoshana Zuboff, The Age of Surveillance Capitalism: the fight for a human future at the new frontier of power, (New York: Profile Books, 2019), p. 8.

74 In the wake of the Snowden revelations, which spoke to government surveillance, people modified their online behaviour to exclude searches on topics they thought might flag them to government. See Jon Penney, Chilling Effects: Online Surveillance and Wikipedia Use, *Berkley Technology Law Journal*, Vol. 31, No. 1, pp. 117-183, 27 April 2016. See also Ahmed Shaheed, Report of the Special Rapporteur, para. 54.

to deal with the issue of targeted advertising under various aspects of its specific legislative framework. At the time of writing, some cases have been adjudicated on and others are in progress.[75] Some of these cases will have implications for protecting autonomy and agency and, therefore, freedom of thought in the new civic space.

> "…the fact that the operator of an online social network holds a dominant position on the market for online social networks does not, as such, preclude the users of such a network from being able validly to consent, (…) to the processing of their personal data by that operator. This is nevertheless an important factor in determining whether the consent was in fact validly and, in particular, freely given, which it is for that operator to prove."
>
> — Press release no. 113/23, Court of Justice of the European Union, Judgment of the Court in Case C-252/2001 | *Meta Platforms and Others* (General terms of use of a social network), Luxembourg, 4 July 2023, para. 154.

Furthermore, academic literature on freedom of thought suggests that this new environment compromises mental autonomy through: (1) the surveillance and use of personal data that gives very detailed insights about individuals or groups; (2) the opacity of technologies, both in how they use AI to generate and propagate content and the inner workings of the technology itself; (3) the exploitation of cognitive biases; and (4) its unprecedented reach, shaping societal preferences and choices in a manner that bypasses democratic processes.[76]

75 Commission finds Apple and Meta in breach of the Digital Markets Act, European Commission Press release, 23 April 2025; see also Press release no. 116/24, Judgment of the Court in Case C-446/21 | *Schrems* (Communication of data to the general public), Court of Justice of the European Union, Luxembourg, 4 October 2024; and Press release no. 113/23, Judgement of the Court in Case C-252/2001 | *Meta Platforms and Others* (General terms of use of a social network), Court of Justice of the European Union, Luxembourg, 4 July 2023. A relevant case was also settled in the UK early in 2025. See Dan Milmo, Meta to stop targeting UK citizen with personalised ads after settling privacy case, *The Guardian,* 22 March 2025.

76 Nina Keese and Mark R. Leiser, Online Manipulation as a Potential Interference with the Right to Freedom of Thought in Patrick O'Callaghan and Bethany Shiner (eds.), *The Cambridge Handbook of the Right to Freedom of Thought*, (Cambridge: Cambridge University Press, 2025). See also Kate Brennan et al, Artificial Power: 2025 Landscape Report.

‘Information disorder’ alters genuine choice

The reorganization of our information ecosystem has also affected the way in which democratic debate and choice operate, with public interest media being severely affected. The proliferation of AI-enhanced digital technologies initiating cycles of disinformation and misinformation for specific political, ideological or commercial gain that spread much faster than before on digital platforms,[77] has produced a state of ‘information disorder’.[78] This phenomenon has damaged trust in independent media and in the integrity of information, and is driving the blurring of truth. As uncertainty increases, public confidence in institutions and the broader information ecosystem is declining, affecting the quality of informed decision-making necessary for democratic participation and in everyday life. [79]

As shown by the OSCE RFoM,[80] GenAI poses additional challenges for journalism. It replaces original content created by journalists, while also creating issues around plagiarism and lack of compensation, all of which could further drive public interest journalism and content out of the public sphere.[81]

In parallel to the shifting role of the media and its ability to act in the public interest by maintaining a healthy information environment,[82] GenAI and the ability to create deepfakes that are increasingly difficult to detect creates further risks for the information landscape and election processes in particular.

---

77 Report of the UNSR on FoOE Irene Khan, Disinformation and Freedom of Opinion and Expression, 13 April 2021, A/HRC/47/25, para. 2.

78 See Claire Wardle PhD and Hossein Derakhshan, *Information Disorder: Towards an interdisciplinary framework for research and policy making*, (Strasbourg: Council of Europe, 2018). Wardle and Derakhshan use a conceptual framework to describe information disorder in the new communication environment, which looks at the types of information disorder (dis-information, mis-information and mal-information); at the three phases of information disorder (creation, production and distribution) and then at the elements of information disorder (agent, message and interpreter). They focus on harm and falseness to define the types of information that make the information disorder ecosystem. Mis-information thus happens when false information is shared, but no harm is intended. Dis-information is when false information is knowingly shared, intending to cause harm. Mal-information is when genuine information is shared intending to cause harm (e.g., making private information public).

79 UNSR on FoOE Irene Khan, Disinformation and Freedom of Opinion and Expression.

80 See Workshop on Big Tech and Media Freedom - Outcome Report, OSCE/RFoM, 15 October 2024.

81 Beyond journalism, GenAI poses significant issues for the respect of Article 15 ICESCR regarding the right of everyone to benefit from the protection of the moral and material interests resulting from any scientific, literary or artistic production of which they are the author.

82 See the work of the OSCE RfOM on the topic: Media and Big Tech Initiative | OSCE. Furthermore, in 2025, the OSCE RFoM is preparing guidance for states on safeguarding media freedom in the age of Big Tech and AI, supporting the development of healthy online information spaces.

These technologies can be easily exploited for manipulation, spreading misinformation and disinformation with devastating effect. Deepfakes are commonly used in gender-based harassment, particularly through the creation and dissemination of sexualized images. These primarily target women and girls, with LGBTI people also among common targets. Women in politics are at particular risk, as well as those living in conservative, patriarchal societies, where a deep fake showing person in an 'inappropriate' situation can have serious consequences.[83]

It has been shown, for example, that malicious actors can use open source GenAI models to perpetrate or amplify gender-based violence (GBV) in the digital space.[84] While GenAI has not created GBV, its power to generate, replicate and disseminate GBV and other types of harmful content at scale and at little cost[85] makes it more destructive if left unregulated and in the hands of such actors, while those affected are pushed out of the online sphere and/or feel too intimidated to react.

GenAI models also present a specific technical challenge, called 'hallucinations'. This is content, produced by GenAI, that moves away from factual reality or is simply fabricated. However, it is usually very plausible and therefore difficult, sometimes impossible, to detect. Hallucinations stem primarily from the way GenAI operates. It is not just problems with the database GenAI works from (including inaccuracies or false information in the case of, for instance, the Internet), but how GenAI is designed. It is meant to 'play along' and produce an answer.[86] OpenAI has recognized that hallucinations, together with other elements, such as intentional misinformation or societal biases, could cast the whole information environment into doubt, threatening our ability to distinguish fact from fiction.[87] This 'losing a sense of reality' may have a real impact on human agency and autonomy, as well as on freedom of thought. While companies and researchers are working

83 Meta Oversight Board, Content Moderation in a New Era for AI and Automation webpage; see also the OSCE/ODIHR CHANGE: Capitalizing on the Human Dimension Mandate to Advance Gender Equality project.

84 Chowdrhury and Lakshmi, "Your opinion doesn't matter, anyway".

85 Laura Bates, Online brothels, sex robots, simulated rape: AI is ushering in a new age of violence against women, *The Guardian,* 3 June 2025.

86 Kate Brennan et al, Artificial Power: 2025 Landscape Report, p. 49.

87 Colleen M Shannon, P. Eng., LL.M., Do AI Hallucinations Disguise Gender Bias?, *Medium,* 27 September 2023.

to reduce hallucinations,[88] including through technical solutions that make sources visible, which is important from many perspectives, it is unlikely that the problem can be solved within the current models and without trade-offs in the performance of the GenAI technology.[89]

The EU AI Act is the first attempt to regulate the use of GenAI, including very sophisticated deepfakes, to protect individuals from AI-facilitated abuse. While some groups are more obviously affected, 'information disorder' has a global impact and places not only freedom of thought, but the human rights and security of all at risk.[90]

The impact of AI on human connection, thought development and the right to be forgotten

The way humans process and understand the world, including through critical thinking, is shaped by their education, upbringing and the conditions for brain development during their formative years. AI, and GenAI in particular, has already had a profound impact on education, and will continue to do so, likely transforming teaching and learning, and at a pace that often leaves little room for appropriate debates on risks and benefits or informed decisions thereafter. Digital and media literacy is essential for all, but curriculums must be updated to equip learners — young and old — with the knowledge and skills to understand, navigate and exert agency over an environment increasingly populated and mediated by AI.

There are potential benefits of AI use in the field of education. However, the use of AI, including GenAI, in education also carries specific risks that must be carefully addressed. Indeed, UNESCO has highlighted potential ethical challenges associated with the adoption of GenAI in educational settings, calling for comprehensive long-term impact assessments.[91] A main concern is the potential reduction in human-to-human interaction. This interaction and human connection is central in the social-emotional aspects of learning, which play an important role in cognitive development and emotional well-being.

88 Reece Rogers, Reduce AI Hallucinations with This Neat Software Trick, *Wired,* 14 June 2024.

89 Kate Brennan et al, Artificial Power: 2025 Landscape Report, p. 49.

90 See also Marina Nord et.al., Media Freedom, Democracy, and Security.

91 Miao and Holmes, Guidance for generative AI in education and research.

There is a risk that the experiences of learning and the acquisition of thinking skills are in danger of being replaced by GenAI-offered solutions. Indeed, academic research on the use of GenAI shows a worrying decline in critical thinking skills.[92] Recent research from industry has similarly shown a decrease in higher order thinking skills among professionals using GenAI models to perform their work.[93] There is also concern that GenAI may narrow exposure to different voices, as it tends to reflect 'standard' answers, further marginalizing under-represented voices.[94]

For AI to have a positive effect in education, integrating GenAI into educational settings requires a clear understanding of its impact on learning and the acquisition of 'thinking skills', followed by clear guidance and training for educators. To be beneficial, it must also be subordinated to educational objectives related to the appropriate development of thinking abilities and human flourishing. In this way, AI could, for instance, enhance personalized learning through adaptive learning platforms, provide access to educational resources in underserved areas and support teachers by reducing redundant tasks such as automated marking of multiple-choice tests, thereby potentially allowing space and opportunity to enrich freedom of thought.

A growing area of concern in new GenAI is the deployment of chat bots designed to simulate human relationships for companionship, including life companions. Young people or people with mental health challenges can be at particular risk. Mental health vulnerabilities may be exacerbated, and terrible consequences occur when chatbots encourage unhealthy behaviours or even advocate suicide.[95] There are particular concerns when AI is used to address mental health issues; the development of appropriate AI-based mental health services is complex, and the risks of using chat bots, including those built for something else (e.g., entertainment or marketing) are significant. Professionals, including psychological associations, have

92 Andrew R Chow, ChatGPT's Impact on Our Brains According to an MIT Study, *Time*, 23 June 2025.

93 Despite plans to invest $80 bn, Microsoft admits that AI is making us dumb, *Business Standard*, 17 February 2025, original source Parmy Olson, *Bloomberg*; Amanda Silberling, Is AI making us dumb?, *TechCrunch,* 10 February 2025.

94 Miao and Holmes, Guidance for generative AI in education and research.

95 Eileen Guo, An AI chatbot told a user how to kill himself — but the company doesn't want to "censor" it, *MIT Technology Review*, 6 February 2025; Kate Payne, In lawsuit over teen's death, judge rejects arguments that AI chatbots have free speech rights, *Associated Press*, 21 May 2025.

warned about such dangers.[96] Immediate regulation, including banning certain applications or uses, is urgently needed to prevent harm and ensure the safe development of any technologies in this field.

The extent to which a person is consciously present and active in public spaces has also changed significantly in the age of AI-powered online spaces. An inherent part of freedom of thought and conscience is our potential, as human beings, to be in constant evolution and change.[97] However, there is now a permanence to everyone's online public presence (e.g., through what they post online), including publicly available facts and the digitization of newspaper archives. For example, past thoughts expressed online can be held against a person today. This can be argued to limit the possibility and freedom to evolve in one's own eyes and in the eyes of others. As such, the 'right to become' that Charles Malik, a drafter of the UDHR, described as the essence of freedom, risks being cut off at its source.

A 'right to be forgotten' is beginning to be recognized in jurisprudence in relation to the right to privacy. This includes the right to have information about oneself that would otherwise remain permanently available in the public sphere erased, except in specific situations, especially where public interest considerations might prevail in terms of having access to certain information.[98]

However, people's digital presence has evolved into what is called a digital footprint. This is a record of all of someone's online activity, not just what they intentionally post or publish for others to see. It includes the websites they visit (recorded through cookies) and what these may say about them. This data makes increasingly sophisticated profiling possible — the extent and results of which one is not aware — affecting individual autonomy. Beyond the basic targeting with tailored content, this profiling could

96 Zara Abrams, Using generic AI chatbots for mental health support: A dangerous trend; American Psychological Association, 12 March 2025; APA calls for guardrails, education, to protect adolescent AI users, American Psychological Association, Press Release, 3 June 2025; Nathalie Koubayová, Meet ChatPal, the European bot against loneliness, Algorithm Watch, 22 May 2023.

97 See, for example, Hannah Arendt, *The Human Condition*, (University of Chicago Press: 1953).

98 "The choice of measure to be implemented in the specific circumstances of each case may vary depending on factors such as the veracity or inaccuracy of the information, the extent to which the information contributes to a debate of public interest, whether it has any historical, research-related or statistical interest, the negative repercussions on the individual's personal sphere of the continued availability of the information online, as well as the amount of time that has elapsed since the events referred in the article, or since the publication of the information." Joint Factsheet: The Right to be Forgotten. ECtHR and CJEU Case-Law, European Court of Human Rights and the EU Agency for Fundamental Rights, 30 October 2024.

become part of wider surveillance systems and could pose cybersecurity threats, or be used to persecute, for example, human rights defenders, dissenters or free thinkers in authoritarian regimes.[99]

99 See the Front Line Defenders webpage on Digital Protection.

5. Neurotechnology and freedom of thought

Another field, highly relevant for freedom of thought, and especially when intertwined with AI, is neurotechnology. This section provides an overview of the key challenges this field poses to freedom of thought.

In a seminal 2024 report,[100] the UN Human Rights Council Advisory Committee refers to "neurotechnology" as encompassing "an array of devices and systems that interact with the central nervous system through electrical, magnetic, optogenetic and other means." Some of these devices and systems are primarily meant to understand the brain's functioning, while others directly intervene in mental processes to restore lost functions and enhance cognitive abilities.

Recent developments in neurotechnology have increased our understanding of the brain and provided new treatments for various neurological or mental health conditions. Large, state-funded research initiatives have contributed to such advances.[101] Nonetheless, the rapid development of neurotechnologies, including their combination with artificial intelligence, raises serious concerns for mental privacy, integrity and autonomy.

In a 2019 recommendation, the OECD Council recognized that "there are ethical, legal and societal questions raised by certain applications of neurotechnologies given the perceived centrality of the brain and cognitive function to notions of human identity, freedom of thought, autonomy, privacy, and human flourishing" and that "a broad public discussion about

100 UN Human Rights Council Advisory Committee, Impact, opportunities and challenges of neurotechnology with regard to the promotion and protection of all human rights, A/HRC.57/61, 8 August 2024, para. 4. Another definition for neurotechnologies is "… devices and procedures used to access, monitor, investigate, assess, manipulate, and/or emulate the structure and function of the neural systems of animals or humans", Preliminary study on the technical and legal aspects relating to the desirability of a standard-setting instrument on the ethics of neurotechnology, UNESCO, 6 April 2023.

101 For example, the US Brain Research through Advancing Innovative Technologies (BRAIN) Initiative, the EU Human Brain Project, or the China Brain project, among others. See also Hermann Garden, David E Winickoff, Nina Maria Frahm, Sebastian Pfotenhauer, Responsible innovation in neurotechnology enterprises, OECD Science, Technology and Industry Working Papers 2019/05, p. 15; or the International Brain Initiative.

the best future of neurotechnology in society" is warranted.[102] In the draft text of a recommendation on the ethics of neurotechnology,[103] UNESCO also considers the ethical, legal and societal issues and questions raised by the application of neurotechnologies with regards to human rights and human dignity including "autonomy, privacy, mental and physical integrity, personal identity, freedom of thought, risk of discrimination, inequality and challenges to democracy (…)." It further recognizes that the sensitivity of interventions in relation to the highly complex nervous system comes from its role in coordinating behaviour and mental processes: "It enables the exercise of individual autonomy, the capacity to act as moral agents, to be responsible for actions, cooperate with others, deliberate about collective decisions, and develop personality."

While neurotechnology has been used in medicine for some time now, its convergence with engineering and computational sciences, alongside the growth of medical and commercial applications, has dramatically changed its impact, advancing medical science, but also introducing risks to freedom of thought and other human rights.[104]

> "… the unprecedented capacity [neurotechnologies] offer to external actors to affect an individual's enjoyment of rights raises enormous ethical questions and challenges the very understanding of the foundational principles of human rights". Neurotechnologies are socially disruptive because they generally: "(a) enable the exposition of cognitive processes; (b) enable the direct alteration of a person's mental processes and thoughts; (c) bypass the individual's conscious control or awareness; (d) enable non-consensual external access to thoughts, emotions and mental states; (e) are nurtured by 'neurodata', which are needed for their own functioning, calibration and optimization; and (f) collect, analyse and process large personal datasets of a highly sensitive nature."
>
> — UN Human Rights Council Advisory Committee, Impact, opportunities and challenges of neurotechnology with regard to the promotion and protection of all human rights, A/HRC.57/61, 8 August 2024.

102 Recommendation of the Council on Responsible Innovation in Neurotechnology, Organisation for Economic Co-operation and Development, OECD/LEGAL/0457, adopted on 11 December 2019.

103 Draft text of the Recommendation on the Ethics of Neurotechnology, UNESCO, 9 April 2025.

104 Human Rights Council Advisory Committee, A/HRC.57/61.

In addition to medical uses, the use of neurotechnology-based products, through apps and devices, gaming experiences, health monitoring, meditation aids or even embedded in earphones, has made brain data accessible to technology companies with consequences for freedom of thought.[105] As UNESCO asserts: "Companies can use neural data obtained from non-invasive neurotech devices for marketing purposes. By detecting signals related to our preferences and dislikes, these companies can influence customer's behaviour for profit maximization. This raises alarming questions about the impact of surveillance, marketing tactics and political influence on our most private thoughts and emotions, ultimately threatening our democracies and the foundations of society."[106] Other sensor technologies can indirectly collect data about our neural activity (e.g., eye tracking, voice recognition and analysis, facial-emotion recognition, etc.) and become problematic when used to infer mental states.[107]

> "It is not simply a question of health that is at stake here, but rather our view of the human person, of our dignity and of the full capacity to exercise our rights in a context of tension between health needs and market aims. On the one hand, we have major health needs, since diseases of the nervous system, neurological diseases and mental illnesses represent one third of our health care expenditure (…) On the other hand, the consumer market, 'neural data' (also called 'brain data') are becoming a sought-after data type and commodity beyond the medical sector including digital phenotyping, affective computing, neurogaming and neuromarketing. Among the tensions raised by neurotechnologies we should also mention public trust, respect for mental privacy, rapid technological and economic development, and the fact that such developments face little or poorly supervised uses."
>
> — Hervé Chneiweiss, Ethics issues and global governance of neurotechnologies, in *The risks and challenges of neurotechnologies for human rights*, UNESCO, University of Milan-Biocca, and State University of New York (SUNY) Downstate, 2023, pp. 48-49.

105 Farahany, *The Battle for Your Brain*.

106 Ethics of neurotechnology, UNESCO webpage.

107 UNESCO, Draft text of the Recommendation on the Ethics of Neurotechnology.

In the aforementioned report, the Human Rights Council Advisory Committee concludes that neurotechnologies affect human rights in a unique manner. The report explores and explains the key risks to human rights, including freedom of thought, posed by the new landscape of neurotechnologies and their (potential) uses, such as the following:

- Technologies initially developed to assist individuals with neurological conditions are now being developed and commercially marketed as apps for 'cognitive enhancement' and other, non-medical purposes. These technologies would allow users to control certain elements of their external environment with their thoughts, or communicate with others who have similar technologies implanted. While the medical solutions they provide can be ground-breaking, the absence of appropriate ethical and human rights frameworks for non-medical devices raises significant concerns, particularly for freedom of thought and non-discrimination, including ensuring equitable access for medical purposes. Concerns about the rights of children come into play, for example, if brain-computer interfaces (that have not been tested for their short or long-term effects on (mental) health) are used in gaming, or where neurotechnology devices might enhance intellectual capabilities.
- Even if full-blown 'thought-reading' is not yet possible, AI-enhanced neurotechnologies, are increasingly capable of making nuanced inferences about thoughts and mental states and allow profiling of individuals in particularly intrusive ways. Specific privacy concerns arise, because neurotechnologies can generate detailed inferences about people's identities (including personality traits, cognitive performance or sexual orientation). Additional risks emerge if this data is used in justice or national security systems, in suspect or witness interrogation, including through neurotechnology-enhanced lie detectors. It could violate the right not to testify against oneself, running the risk of individuals being punished for their thoughts. This includes situations where the technology may still be too inaccurate to be correct about the thoughts in question. Revealing unexpressed thought without the consent of the 'owner' runs counter to the essence of freedom of thought.[108]
- Devices for monitoring mental states are already in use in work environments, particularly under extreme conditions, to ensure vigilance and

108 Human Rights Council Advisory Committee, A/HRC.57/61.

avoid accidents due to fatigue.[109] While the stakes may be high, their deployment raises serious human rights concerns, including as related to worker's rights and freedom of thought. This could open the door to neurosurveillance for productivity purposes.[110] The EU AI Act took steps to address this issue for specific environments.[111] More generally, the consumer application of these technologies, given their unknown long-term effects and particularly if deployed without clear ethical and human rights oversight, poses risks to both personal integrity and (mental) health.

- Groups in vulnerable situations, such as the elderly, children, people with disabilities and people deprived of liberty or in other custodial settings, are especially at risk while there is no proper regulation of the development and use of neurotechnologies. There are particular issues over consent, which should always be prior, free, informed, real, transparent, effective and never assumed. Given the medical purposes for which many of these technologies are developed, people with disabilities and their representatives, for example, should be included in the development process, with their needs, rights and perspectives as end users prioritized. Respecting the rights of people with disabilities also includes ensuring they have full access to the neurotechnologies once developed, and that they are safe, effective and respect human rights in their design, development and use.[112] This approach also aligns with the social model of the Convention on the Rights of Persons with Disabilities, which views disability as part of human diversity and society, as opposed to the ableist medical model that historically focused on prevention and cure.[113]

Regulations should aim to protect human rights, by defining clear human rights-compliant frameworks for commercial uses, while keeping the achievement of medical breakthroughs as a priority.

---

109 These include fields such as mining, construction, trucking, aviation, railways. Farahany, *The Battle for Your Brain*. See also José M. Muñoz, Laura Isaza, and Tarini Mehta, Tech is coming for your brain data: how a Chilean politician turbocharged the "neurorights" movement, *The Boston Globe*, 10 September 2024.

110 Human Rights Council Advisory Committee, A/HRC.57/61.

111 High-level summary of the AI Act, EU Artificial Intelligence Act, 27 February 2024, updated on 30 May 2024. Under the Act, inferring emotions in workplaces or educational institutions, except for medical or safety reasons, falls under prohibited AI systems. See also Nora Santalu, Neurotechnologies under the EU AI Act: Where law meets science, IAPP, 12 May, 2025.

112 Human Rights Council Advisory Committee, A/HRC.57/61.

113 International Disability Alliance, *Submission to the Human Rights Committee Advisory Committee call for inputs on neurotechnology and human rights*, 2 July 2023, see the OHCHR Neurotechnology and human rights webpage.

6. Regulatory frameworks

The regulation of AI throughout its lifecycle has lagged behind in one of the fastest growing and most complex and consequential fields in the history of humanity. While various non-binding documents exist at regional or global level, there is no single, universal, legally-binding standard on AI that protects human rights.[114]

> In adopting its Recommendation on the Ethics of AI in 2021, the UNESCO General Conference did so in recognition "of the profound and dynamic positive and negative impacts of artificial intelligence (AI) on societies, environment, ecosystems and human lives, including the human mind, in part because of the new ways in which its use influences human thinking, interaction and decision-making and affects education, human, social, and natural sciences, culture, and communication and information." The Recommendation primarily aims to protect human rights and dignity, based on a set of principles, such as transparency, fairness or human oversight. In its preamble, it also warns of various asymmetries of power around AI, which raise diverse and serious risks and potential consequences, and advocates for stronger global cooperation and solidarity, including multilateralism.
>
> — Recommendation on the Ethics of Artificial Intelligence, UNESCO, 23 November 2024

By regulating the field through the EU AI Act, the EU has made significant progress, including on human rights protections. Since it has only been adopted recently and needs further guidance and instrument development, as well as national-level implementation frameworks, the Act has yet to demonstrate its capacity to mitigate human rights impacts. In 2024, the Council of Europe also adopted a Framework Convention on Artificial

114 Artificial Intelligence, UNESCO webpage; UNESCO, Draft text of the Recommendation on the Ethics of Neurotechnology; Artificial intelligence, OECD webpage; Governing AI for Humanity: Final Report, United Nations AI Advisory Body, September 2024; UN Global Digital Compact.

Intelligence and Human Rights, Democracy and the Rule of Law, which includes a larger number of countries under its framework.[115]

While concerns have also been raised about both instruments in terms of scope and effectiveness of protections,[116] it is important that instruments such as the Framework Convention are ratified, fully implemented and strengthened where needed. This requires resources and commitment at state level and from EU institutions in the case of the EU AI Act. While non-binding, and not a substitute for regulation, industry standards also contribute to creating ethical frameworks.[117] ODIHR also notes the emergence of academic work and initiatives that specifically look at the need for global regulatory frameworks and inform debates on this topic.[118]

The EU AI Act is currently the most comprehensive attempt to create a framework regulating AI development and use, far ahead of other regulatory frameworks. It adopts a risk-based approach, categorizing AI systems into prohibited, high-risk, limited risk and minimal risk. It regulates general purpose AI — essentially those operating on foundational models, such as GenAI — and imposes a number of obligations, including on transparency and accountability. It also defines general purpose AI that poses systemic risks (i.e., powerful models), imposing further requirements and clarifying that such systems can be used as high-risk AI systems, or integrated into them, which brings specific obligations to cooperate with such high-risk systems to enable their compliance under the Act. Importantly, free and open licence general purpose AI models have to comply with fewer

115 It is also open for signature to the non-Member States of the Council of Europe that helped develop it, and there are specific, separate procedures allowing other, non-Member States to join. As of January 2025, it had been signed by the United States and Israel, outside the Council of Europe. For it to be binding, it also requires ratification.

116 See EU: Artificial Intelligence rulebook fails to stop proliferation of abusive technologies, Amnesty International, 13 March 2024; and ENNHRI Calls on Council of Europe member States to ensure strong human rights protection in the draft Convention on AI, Human Rights, Democracy and Rule of Law, ENNHRI, 17 May 2024. For a comprehensive analysis of the Act, as well as other related legislation, see Sandra Wachter, Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This means for the European Union, the United States, and Beyond*, Yale Journal of Law and Technology*, Vol. 26, Issue 3, pp. 671-718; for a discussion of specific neurotechnologies, including gaps in EU legal coverage, see Christoph Bublitz, Banning biometric mind reading: the case for criminalizing mind probing, *Law, Innovation and Technology*, Vol. 16, Issue 2, pp. 432–462.

117 See, for example, IEEE Frameworks, Institute of Electrical and Electronics Engineers (IEEE) Tech Ethics.

118 See, for example, Alexander Kriebitz, Caitlin C Corrigan (eds.), Promoting and Advancing Human Rights in Global AI Ecosystems: The Need for A Comprehensive Framework under International Law, February 2025.

obligations under the EU AI Act.[119] Risk-based frameworks like the EU AI Act can serve as models for other regions.

> Prohibited AI systems include those that: deploy subliminal, manipulative or deceptive techniques to change behaviour and hamper informed decision-making, and that cause significant harm; exploit people's vulnerabilities (e.g., age, disability, socio-economic circumstances) to distort behaviour, and causing significant harm; biometric categorization systems (with exceptions for law enforcement); perform social scoring based on social behaviour or personality traits; assess the risk of an individual criminally offending solely based on profiling or personality traits, except when used to enhance human assessments based on objective, verifiable facts directly linked to criminal activity; compile facial recognition databases by indiscriminate harvesting of facial images from the internet or CCTV; infer emotions in workplaces or educational institutions, except for situations where medical or safety concerns arise; perform remote biometric identification in real time and in public spaces for law enforcement purposes, with some exceptions that require, among others, a fundamental rights impact assessment.
>
> High-risk AI systems providers are subject to specific requirements, including risk management, data quality, transparency, accountability and human oversight. Before deployment, with some exceptions and under certain conditions, these systems also have to undergo a fundamental rights impact assessment. AI systems that profile individuals are always considered high-risk.
>
> — Drawn from the High-level summary of the AI Act, EU Artificial Intelligence Act webpage, 27 February 2024, last updated on 30 May 2024.

Before the AI Act, the EU adopted the Digital Services Act (DSA) package, which regulates online platforms and intermediaries, with specific rules for very large platforms and search engines. The DSA also focuses on addressing systemic risks including in the areas of: illegal content, freedom

119 High-level summary of the AI Act, EU Artificial Intelligence Act, 27 February 2024, updated on 30 May 2024.

of expression, media freedom and pluralism, discrimination, consumer protection and children's rights, public security and electoral processes, gender-based violence, and mental and physical well-being. It also requires a certain level of transparency and oversight by authorities and an option for recommender systems not based on user profiling.[120]

This EU regulatory framework operates in an extremely complex field, and its impact is only starting to be seen and understood. The necessary instruments and institutional frameworks, especially at the national level, are still being developed/established at the time of writing.[121] Furthermore, the move towards securing human rights protections in relation to AI through regulation is not a given.[122] While important, the EU framework does not apply to the whole OSCE region, let alone the world, and global regulation is necessary.[123] Both the EU AI Act and the Council of Europe Framework Convention include exceptions for national security purposes. Given that the EU is the most advanced regulator of AI in the world, the national security exceptions raise significant human rights concerns, potentially contributing to a new global arms race, where AI supremacy is seen as the key to dominance.[124] To mitigate this, international cooperation is essential to develop harmonized standards that close such gaps and prevent misuse while promoting responsible innovation.

While the development of products in the medical field is subject to certain regulations, neurotechnologies that are developed for commercial purposes raise new risks for human rights and fundamental freedoms. These technologies require stronger frameworks for protection. Noting that there are certain, but generally limited, elements of regulation at national or regional[125]

120 European Commission, DSA: Very large online platforms and search engines (webpage last updated 12 February 2025); and European Commission, The Digital Services Act package (webpage last updates 12 February 2025).

121 See Towards meaningful fundamental rights impact assessments under the DSA, European Center for Not-for-Profit Law and Access Now, 15 September 2023; and Aninda Chakraborty, LatticeFlow AI unveils EU compliance framework for Generative AI, *Tech Monitor*, 17 October 2024; Overview of all AI Act National Implementation Plans, EU Artificial Intelligence Act website, 8 November 2024 (last update 19 May 2025).

122 Withdrawal of the AI Liability Directive Proposal Raises Concerns Over Justice for AI Victims, Center for Democracy and Technology, 12 February 2025.

123 Some states outside the OSCE region have adopted various regulatory frameworks, but they are not the subject of this analysis. The EU remains the most advanced and complex regulator in this field.

124 Ilaria Carroza, Nicholas Marsh and Gregory M. Reichberg, Dual-Use AI Technology in China, the US and the EU: Strategic Implications for the Balance of Power. PRIO Paper, (Oslo: PRIO, 2022).

125 See, for example, OECD, Recommendation of the Council on Responsible Innovation in Neurotechnology. OECD/Legal/0457, adopted on 11 December 2019.

levels or within the EU AI Act, there are no universal, international, legally-binding standards on neurotechnologies. At UNESCO, Member States are currently negotiating the first, non-binding, global, standard-setting instrument on the ethics of neurotechnology, following from its earlier work on the ethics of AI.[126] Beyond this, comprehensive regulation is required that will cover the full complexity of the new challenges neurotechnologies pose, both through existing protection frameworks that could be expanded to encompass the specific new risks to human rights, and through new instruments on the development and use of neurotechnologies.

In a 2024 report,[127] the UN Special Rapporteur on the Right to Privacy proposed text for a resolution updating the 1990 General Assembly Resolution 45/95, "Guidelines for the regulation of computerized data files". The proposed text states that neurodata processing "must not be used to manipulate or alter the freedom of thought and consciousness of an individual, making him or her dependent on a third party or altering his or her ideas, security or independence or his or her natural cerebral identity or neurocognitive integrity." Furthermore, such data may not "be processed for purposes other than the promotion of health and the diagnosis, rehabilitation and alleviation of disease in the context of the right to health, or scientific research in the fields of biology, psychology and medicine aimed at alleviating suffering or improving health." In April 2025, through Resolution 58/6, the UN Human Rights Council recognized that "the continued development of some of its applications may pose a number of ethical, legal and societal questions and has implications for human dignity and autonomy, making it necessary to ensure that human rights are effectively respected, protected and fulfilled in this context". It requested the Advisory Committee to draft a set of guidelines for applying the existing human rights framework to the conception, design, development, testing, use and deployment of neurotechnologies.[128]

Although many complexities persist and the implications are still unclear, people's understanding of how social media uses their data has rapidly advanced in recent years. By contrast, their understanding of the human

126 The Ethics of Neurotechnology: UNESCO appoints international expert group to prepare a new global standard, UNESCO Press release, 22 April 2024.

127 A/79/173: Report of the Special Rapporteur on the right to privacy, Ana Brian Nougrères - Proposal for the updating of General Assembly resolution 45/95 of 14 December 1990, entitled "Guidelines for the regulation of computerized personal data files", OHCHR, 17 July 2024.

128 Human rights guidelines on neurotechnology, UN Human Rights Council webpage.

rights implications of giving up their neurodata remains extremely low, especially given the lack of vvan appropriate regulatory framework.[129]

The 'AI revolution' earns its name from the profound way it has transformed human life and will continue to do so. However, this revolution has largely taken place without any public debate or contribution to decision-making. Having consumers operating in the digital market, only partially aware and in limited control of their choices, is not the same as democratic deliberation. Likewise, the rapid development of neurotechnologies, and in particular the commodification of devices that interact with the human brain or neural circuits, have an impact on human rights in an unprecedented and profound manner. OSCE participating States have a duty to reinstate the conditions necessary for public debate on issues of critical societal importance. Issues such as the 'cognitive enhancement' of healthy individuals, for example, or possibly even 'human augmentation' bear such profound implications for humanity that they deserve genuine and thorough public debate to inform decisions on how to regulate such technologies. Debates and decision-making must include a broader range of people than the tiny number engaged in developing the technology or the few engaged in research who can claim to have an informed opinion on the subject.[130] Even if the topics are complex, their implications for society should be carefully distilled into clearly understandable information and shared with the general public, so that genuine democratic deliberation can happen, including on whether to use AI or not.

> "The AI hype (…) has sucked the air out of an already stuffy room, making it feel futile—at times impossible—to imagine anything other than a steady march toward the inevitable supremacy of AI. But no matter how true that may feel, it is only that: a feeling. It is not reality—not yet, at least. There are, in fact, many alternatives to this version of AI, many ways to shape new worlds. Like AI, though, these are not inevitable either. Making them possible starts by asking and answering a single question: Is this the world we want?"
>
> — Kate Brennan, Amba Kak and Dr. Sarah Myers West, *Artificial Power: 2025 Landscape Report*, AI Now Institute, 3 June 2025, pp. 9 and 12.

129 Farahany, *The Battle for Your Brain*.

130 See also OECD, Recommendation of the Council on Responsible Innovation in Neurotechnology, Recommendation No. 5: Enable societal deliberation on neurotechnology.

7. Conclusions and recommendations

The need to protect the right to freedom of thought is becoming increasingly urgent, given the extent to which AI-based technologies are shaping both societies and individuals' sense of self. The current framework for the development and use of these technologies deepens power imbalances and poses risks to democracy. Thought and decision-making must be protected against novel forms of subtle but effective manipulation, while unexpressed thoughts must remain private. In the absence of adequate regulatory frameworks to protect human rights, the proliferation of AI-based technologies seems to be blurring, if not erasing, the contours of democratic responsibility and accountability. States have a duty to respect, protect and fulfil human rights, which requires specific regulation at national and international levels. States may be late in devising regulation, but there is still time for them to do so, in particular for the development and use of neurotechnologies to mitigate the risks of the widespread harm they may do.

Building on a set of non-exhaustive criteria proposed by the UN Special Rapporteur on freedom of religion or belief to assess, on a case-by-case basis, what may amount to impermissible manipulation of thought, Susie Alegre and Aaron Schull outline a set of factors for a legal test to assess technology-enabled, unlawful manipulation. They note that the right to freedom of thought, as part of the general human rights framework under the UN, "evolved in response to the atrocities committed by Nazi Germany and in recognition of the unfathomable risks of the brainwashing of populations to human rights more broadly. The lines around effective protection of the right to freedom of thought are crucial for the future of democracy and human rights around the world." They go on to propose that "[a] new legal test should be set out under a General Comment by the HRC that takes account of a range of additional factors:

- The scope and scale of application of a practice;
- Whether the general public or individuals concerned are aware of a tactic for influence;

- Whether the general public or individuals concerned can understand the impact of a particular practice;
- Whether the practice is designed to bypass rational faculties;
- The intensity and period of time of exposure;
- The targeting of particular cognitive biases or vulnerabilities (especially when based on "insider knowledge" drawn from data);
- Any power imbalance; and
- He practical ability to say no."

— Susie Alegre and Aaron Schull, *Freedom of Thought: Reviving and Protecting a Forgotten Human Rights,* Center for International Governance Innovation, 2024, p. 12.

In light of the arguments presented in this document and to create an enabling environment for freedom of thought, OSCE participating States are recommended to:

- **Adopt human rights-based policies, legislation and regulations at the national level and establish institutional infrastructure** to ensure the respect, protection and fulfilment of the right to freedom of thought, and all human rights and fundamental freedoms, in the context of the growing role of AI-based technologies/that use AI, including neurotechnologies. Additionally, to ensure respect for human dignity and democratic principles, it will be critical to regulate neurotechnologies strictly based on ethical and human rights considerations, including by avoiding the social media business model, while incorporating risk-based approaches and ensuring human rights protections across the AI lifecycle as developed by industry.
- **Adopt human rights impact assessment methodologies** as standard practice, building the legal and institutional framework necessary to ensure the widespread use of such methodologies. These tools must prioritize the respect, protection and fulfilment of human rights throughout the AI life cycle and, through regular review, adapt to the rapid pace of AI development and new human rights challenges it brings. **Develop tools to integrate freedom of thought into human rights impact assessments** for all AI-based technologies, considering it both as a separate right and as interdependent with other rights, particularly freedom of expression and the right to a private life.

- **Collect data on AI throughout its life cycle, looking at compliance with human rights standards (including freedom of thought), potential good practices and ways to benefit from AI, as well as assessing access and equity.** Given the rapid uptake of new, AI-based technologies in different fields, the world is falling behind not only in regulation, but also in the effective monitoring and mapping needed to understand the effects on human rights, anticipate implications and identify emerging opportunities. States have a duty to monitor and ensure respect for human rights. Particular care is warranted for the most vulnerable, including people with disabilities, children, the elderly or those in places of deprivation of liberty.
- **Foster and sustain new research and conceptual clarification on standards** for the promotion and protection of freedom of thought, especially in the context of new, AI-based technologies. This should include international cooperation in UN forums, (e.g., through the adoption of Human Rights Council resolutions), as well as activating UN human rights mechanisms (e.g., requesting General Comments/reports) or through other, similar means.
- **Work with and prepare education systems** to reflect new modes of social life and help learners understand and act as critical thinkers, with autonomy and agency, in order to be able to preserve their freedom of thought in a world increasingly powered by AI. This will probably require more than today's digital literacy efforts. New national curriculums should cover the content, skills and behaviours necessary to prepare current and future generations for these changes, which are affecting all areas of life. Adult learning and awareness-raising campaigns should also be implemented.
- **Work with all educational and training institutions to provide ethics and human rights training for those working or deciding on the use of AI throughout its life cycle to ensure it is built to respect and, ideally, enhance the protection and fulfilment of human rights.** Training while studying (e.g., at university) and continuous training for those already working should ensure that those developing, deploying or operating AI and new technologies (from data scientists to engineers and others) have sufficient training in the ethical and human rights implications of the technology, including as related to freedom of thought. Professionals working with AI throughout its lifecycle (including those making decisions over purchase and deployment) should know

how to conduct human rights impact assessments and embed human rights protections into AI-technologies by design.

- **Protect independent media and strengthen its public interest role, also in protecting freedom of thought, freedom of opinion and freedom of expression, including by implementing the** OSCE/RFoM recommendations on this highly important topic.[131]
- **Foster genuine public debate on the ethical and human rights implications, including on freedom of thought, of AI and neuro-technologies.** These should be structured and multidisciplinary, and engage diverse stakeholders, especially those whose rights may be affected or end users of the technologies. The debates should involve public interest media to raise public awareness on these implications, and ensure information is distilled to facilitate an informed public debate. They should be transparent and accountable, with the results reflected in policy decisions.
- **Work with the OSCE's independent institutions, such as ODIHR and the RFoM, to ensure the regulation and implementation of a human rights-compliant AI life cycle**.

131 For more information, see the OSCE RFoM Spotlight on Artificial intelligence and Freedom of Expression — SAIFE Resource Hub and the Policy manual in particular as well as upcoming work on Media and Big Tech.

Office for Democratic Institutions and Human Rights

OSCE ODIHR