



# UNLAWFUL BY DESIGN

EXPOSING THE HUMAN RIGHTS COSTS  
OF GENERATIVE AI



**Amnesty International is a movement of 10 million people which mobilizes the humanity in everyone and campaigns for change so we can all enjoy our human rights. Our vision is of a world where those in power keep their promises, respect international law and are held to account. We are independent of any government, political ideology, economic interest or religion and are funded mainly by our membership and individual donations. We believe that acting in solidarity and compassion with people everywhere can change our societies for the better.**

© Amnesty International 2026

Except where otherwise noted, content in this document is licensed under a Creative Commons (attribution, non-commercial, no derivatives, international 4.0) licence.

<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

For more information please visit the permissions page on our website: [www.amnesty.org](http://www.amnesty.org)

Where material is attributed to a copyright owner other than Amnesty International this material is not subject to the Creative Commons licence.

First published in 2026

by Amnesty International Ltd

Peter Benenson House, 1 Easton Street

London WC1X 0DW, UK

Index: POL 40/0996/2026

Original language: English

**amnesty.org**



*Cover illustration: Kassý Cho*

**AMNESTY**  
INTERNATIONAL



# CONTENTS

<b>1. EXECUTIVE SUMMARY</b>	<b>7</b>
<b>2. TECHNICAL BACKGROUND TO GENERATIVE AI</b>	<b>11</b>
2.1 LARGE LANGUAGE MODELS (LLMs)	12
2.2 GENERATIVE ADVERSARIAL MODELS (GANs)	12
2.3 VARIATIONAL AUTOENCODERS (VAEs)	12
2.4 DIFFUSION MODELS	13
2.5 STANDALONE GENERATIVE AI TOOLS	13
<b>3. RELEVANT LEGAL STANDARDS</b>	<b>14</b>
3.1 THE RIGHT TO PRIVACY	14
3.2 THE RIGHT TO EQUALITY AND NON-DISCRIMINATION	15
3.3 THE RIGHT TO FREEDOM OF EXPRESSION	17
3.4 FREEDOM OF THOUGHT	18
3.5 BUSINESS AND HUMAN RIGHTS STANDARDS	19
<b>4. BACKGROUND TO GLOBAL DEVELOPMENTS IN GENERATIVE AI</b>	<b>21</b>
<b>5. HUMAN RIGHTS TENSIONS IN GENERATIVE AI DESIGN</b>	<b>25</b>
5.1 UNLAWFUL WEB-SCRAPING AND MASS INVASIONS OF PRIVACY BY DESIGN	25
5.2 DATA CENTRES AND ENVIRONMENTAL IMPACT	28
5.3 PERPETUATING STEREOTYPES, BIASES AND DISCRIMINATION	30
5.4 SHRINKING CIVIC SPACE	32
5.5 AUTOMATION BIAS AND MANIPULATION	34
5.6 CONCLUSION	35
<b>6. HUMAN RIGHTS INCOMPATIBILITY OF CURRENT GENERATIVE AI SYSTEMS BASED ON DATA PIPELINE</b>	<b>37</b>
6.1 THE RIGHT TO PRIVACY	37
6.2 THE RIGHT TO EQUALITY AND NON-DISCRIMINATION	38
6.3 RIGHT TO FREEDOM OF EXPRESSION	38
6.4 RIGHT TO FREEDOM OF THOUGHT	39

6.5 BUSINESSES' HUMAN RIGHTS RESPONSIBILITIES	39
<b>7. CONCLUSIONS AND RECOMMENDATIONS</b>	<b>40</b>
7.1 CONCLUSION	40
7.2 RECOMMENDATIONS	40
TO STATES	40
TO COMPANIES	41
TO UN BODIES AND REGIONAL ACTORS	41

# GLOSSARY

WORD	DESCRIPTION
<b>AUTOMATION BIAS</b>	The tendency for humans to favour suggestions from automated decision-making systems over their own judgement.
<b>CEDAW</b>	Convention on the Elimination of All Forms of Discrimination Against Women
<b>COMMON CRAWL</b>	A large repository of web-scraped data used for training AI models, containing data from millions of web domains.
<b>DARK PATTERNS</b>	Deceptive user interface designs that manipulate users into making certain choices or taking certain actions.
<b>DATA LABELLING</b>	The process of annotating or tagging training data to help AI models learn patterns and relationships.
<b>DIFFUSION MODELS</b>	AI models that generate new data by learning to reverse a gradual noise-addition process, commonly used in image generation.
<b>DSA</b>	EU Digital Services Act
<b>DWP</b>	UK Department of Work and Pensions
<b>EU AI ACT</b>	EU Artificial Intelligence Act
<b>FOUNDATION MODEL</b>	A broad AI model trained on vast amounts of data that can be adapted for various tasks such as image, video, text and sound generation, and complex calculations, also known as general-purpose AI.
<b>FRT</b>	Facial recognition technology
<b>GDPR</b>	General Data Protection Regulation
<b>GENERATIVE ADVERSARIAL NETWORKS (GANS)</b>	A model architecture where two neural networks compete – a generator that creates synthetic data and a discriminator that tries to identify whether data is real or synthetic.
<b>GENERATIVE AI</b>	A type of artificial intelligence that uses machine learning models to produce, or “generate”, new data or content by mimicking the input data on which they have been trained. Generative AI can generate text, images, audio and other media types.
<b>GRAPHICS PROCESSING UNIT (GPU)</b>	Specialized computer chips crucial for running AI models, particularly in data centres.

WORD	DESCRIPTION
HRC	UN Human Rights Committee
ICCPR	International Covenant on Civil and Political Rights
ICERD	International Convention on the Elimination of All Forms of Racial Discrimination
ICERD COMMITTEE	UN Committee on the Elimination of Racial Discrimination
ICESCR	International Covenant on Economic, Social and Cultural Rights
IHRL	International Human Rights Law
ILO	International Labour Organisation
LARGE LANGUAGE MODELS (LLMs)	Machine learning models trained on vast amounts of text data to process and generate human-like language. They serve as foundation models capable of performing multiple tasks like conversation, content generation and text analysis.
OECD GUIDELINES	OECD Guidelines for Multinational Enterprises
OHCHR	UN Office of the High Commissioner for Human Rights
PARAMETERS	The variables within an AI model that are adjusted during training to determine how the model processes input and generates output.
STANDALONE GENERATIVE AI SYSTEMS	Products that are developed, deployed and marketed for their generative AI capabilities solely and specifically, such as AI chatbots, image/video/audio/text generators, and so on. This does not include products where generative AI is an added feature or function in a larger suite of products, for example, word processing software with optional generative AI features. Standalone generative AI products are, in other words, generative AI models that come with their own front-end for direct use that is fundamentally concerned with generating outputs.
SYNTHETIC DATA	Artificially generated data that mimics the characteristics of real data, created by AI models rather than collected from real-world sources.
TRAINING DATA	The dataset used to teach AI models patterns and relationships, which they then use to generate new content or make predictions.
UDHR	Universal Declaration of Human Rights
UN GUIDING PRINCIPLES	UN Guiding Principles on Business and Human Rights
VARIATIONAL AUTOENCODERS (VAES)	Generative AI models that learn to encode input data into a compressed form, thereby obfuscating inputs into seemingly unrecognisable forms, and then decoding it to create variations of the original input.
WEB SCRAPING	Automated process of extracting data from websites, used to collect training data for AI models.

# 1. EXECUTIVE SUMMARY

The dawn of generative artificial intelligence (AI) systems has captured the world by storm, as a technology that often appears – at least at first glance – efficient, sophisticated and capable of carrying out complex human tasks. Companies have been particularly successful at galvanizing fascination and enthusiastic investment in such technologies.

Behind the veneer of efficiency, sophistication and complexity, however, hides a reality of human rights-violating design principles, which are akin to those found in many of the most problematic AI tools predating generative AI.

Generative AI systems are machine learning tools that have been trained on massive datasets – often without the knowledge and consent of those from whom the data originates, such as social media users and artists – to create “new” data or content through mimicking or approximating the data they have been fed. This data often comes in the form of text, images, video, music and other multimodal outputs, and could not exist without the input, processing and transformation of existing personal, creative and behavioural data. Generative AI systems are “prompted” to generate such outputs by a user requesting outputs within particular parameters.

Amnesty International adopts a supply chain lens in its human rights analyses of AI systems, which implies understanding AI products as reliant on complex supply chains that are each comprised of a network of actors responsible for the various aspects of the system’s training, development and deployment.<sup>1</sup> These actors, from graphics processing unit (GPU) manufacturers, data centre providers, web crawlers, model developers and data annotators, to end user design and outputs, co-produce the technology’s functionality.<sup>2</sup> This briefing focuses on the “data pipeline” aspect of the supply chain of generative AI products, specifically, the stages related to data capture, analysis, and processing. It examines the key parameters and implications of design choices concerning the training data of generative AI models, with a focus on methods and sources of data collection, data processing, model scaling and data outputs.

This briefing examines how standalone generative AI systems, based on unlawful web scraping, are in conflict with international human rights law (IHRL) and standards through their design, development and deployment. While these technologies promise sophisticated automation and efficiency, they rely on data collection and model training practices that abuse privacy rights, enable discrimination, and threaten freedom of expression and thought.

Amnesty International defines standalone generative AI tools as products that are developed, deployed and marketed for their generative AI capabilities solely and specifically, such as AI chatbots, image/video/audio/text generators, and so on. This does not include products where generative AI is an added feature or function in a larger suite of products, for example, word processing software with optional generative AI features. Standalone generative AI products are, in other words, generative AI models that come with their own front-end for direct use that is fundamentally concerned with generating outputs.

---

<sup>1</sup> Ian Brown, Allocating accountability in AI supply chains, 29 June 2023, <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/> (accessed on 14 May 2026).

<sup>2</sup> Jennifer Cobbe, Michael Veale, and Jatinder Singh, Understanding accountability in algorithmic supply chains, 12 June 2023, <https://dl.acm.org/doi/10.1145/3593013.3594073> (accessed on 14 May 2026).

We compared popular generative AI tools and their technical specifications against relevant legal standards under IHRL through an analysis of academic scholarship,<sup>3</sup> research reports by civil society organizations,<sup>4</sup> investigative media reports documenting their harms,<sup>5</sup> and product documentation and policies where these were available. This analysis also drew on Amnesty's previous research on artificial intelligence systems. Amnesty International's human rights analysis of generative AI systems, based on the above methodology, finds that **standalone generative AI systems based on unlawful web scraping are rooted in mass invasions of privacy by design and are therefore incompatible with the right to privacy. The large-scale data-scraping and training required to build many generative AI systems have a number of human rights consequences across the wider supply chain and in the downstream use of these tools. These include:**

- As generative AI models have increased in scale and their deployment in commercial and public contexts have expanded, so has the amount of data required to train them. This, in turn, has meant that the infrastructural requirements and associated environmental costs of generative AI models have also increased, as the growing processing needs of larger models requires more energy-intensive chips, larger data centres, and as a result, more energy and water for its operationalisation. This has had disproportionate negative impacts on communities in Global Majority countries, where many of these data centres are increasingly located.
- Generative AI outputs demonstrate systematic biases that discriminate against marginalized groups and amplify existing inequalities, stemming from biases contained in training data. As datasets powering AI models scale up, the presence of hateful and discriminatory content also increases, along with negative stereotypes and prejudices, especially along racial and gendered lines. Research shows consistent racial, gender and cultural biases in system outputs, which reflect, amplify and reinforce and obfuscate discriminatory patterns in the training data. The predominantly English-language training data means that the resourcedness of generative AI models skews towards English, which leads to biases in the outputs that reflect the social, cultural, linguistic and political norms most present in datasets, reinforcing Western and anglophone cultural and linguistic dominance, while discriminating against and disregarding majority world languages, cultures, ideas and representation. The generation of synthetic data based on real individuals' likenesses also violates the right to equality and non-discrimination, especially it perpetuates violence against women and children, with AI-generated child sexual abuse material (CSAM) overwhelmingly portraying girls.<sup>6</sup>
- In areas of rapid adoption and use, such as content moderation on social media platforms, generative AI systems demonstrate particularly heightened risk of overbroad censorship and disproportionately affect historically marginalized communities. When used to support content moderation, LLM-based generative AI tools risk flagging, removing, suppressing or otherwise censoring content in Global Majority languages and differing cultural contexts, through inaccurate and problematic mistranslations, without justification.<sup>7</sup>

<sup>3</sup> For example, Bender, E.M. and others (2021) "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '21), pp. 610–623. Available at: <https://doi.org/10.1145/3442188.3445922>; Belcak, P. and others (2025) 'Small Language Models are the Future of Agentic AI'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2506.02153>; Geoffrey Currie and others, "Gender and ethnicity bias in generative artificial intelligence text-to-image depiction of pharmacists", December 2024, *International Journal of Pharmacy Practice*, Volume 32, Issue 6, <https://doi.org/10.1093/ijpp/riae049>

Ananya, "AI image generators often give racist and sexist results: can they be fixed?", 19 March 2024, *Nature*, Volume 627, Issue 8005, <https://doi.org/10.1038/d41586-024-00674-9>; Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, Multimodal datasets: misogyny, pornography, and malignant stereotypes, 5 October 2021, <http://arxiv.org/abs/2110.01963> (accessed on 14 May 2026)

<sup>4</sup> Ian Brown, Allocating accountability in AI supply chains, 29 June 2023, <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/> (accessed on 14 May 2026); Marissa Newman, "Palantir allegedly supplying Israel with AI tools amid Israel's war in Gaza", 10 January 2024, *Business & Human Rights Resource Centre*, <https://www.business-humanrights.org/en/latest-news/palantir-allegedly-supplying-israel-with-ai-tools-amid-israels-war-in-gaza/>; Big Tech lobbying is derailing the AI Act | Corporate Europe Observatory, <https://corporateeurope.org/en/2023/11/big-tech-lobbying-derailing-ai-act> (accessed on 14 May 2026).

<sup>5</sup> Bobby Allyn, "Microsoft's new AI chatbot has been saying some 'crazy and unhinged things'", 2 March 2023, NPR, <https://www.npr.org/2023/03/02/1159895892/ai-microsoft-bing-chatbot>; Kate Conger and John Yoon, "Explicit deepfake images of Taylor Swift elude safeguards and swamp social media", 26 January 2024, *New York Times*, <https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html>; Ameera Kawash, "What a cat in a keffiyeh reveals about AI's anti-Palestinian bias", 25 April 2023, +972 Magazine, <https://www.972mag.com/ai-bias-palestinian-cat-keffiyeh/>

Liz O'Sullivan and John P. Dickerson, "Here are a few ways GPT-3 can go wrong", 7 August 2020, *TechCrunch*, <https://techcrunch.com/2020/08/07/here-are-a-few-ways-gpt-3-can-go-wrong/>; Ernestas Naprys, "Meta leached 82 terabytes of pirated books to train its Llama AI, documents reveal", 7 February 2025, *Cybernews*, <https://cybernews.com/tech/meta-leached-82-terabytes-of-pirated-books-to-train-its-llama-ai-documents-reveal/>

<sup>6</sup> Internet Watch Foundation, What has changed in the AI CSAM landscape?, July 2024, p. 23, [https://www.iwf.org.uk/media/opkpmx5q/iwf-ai-csam-report\\_update-public-jul24v11.pdf](https://www.iwf.org.uk/media/opkpmx5q/iwf-ai-csam-report_update-public-jul24v11.pdf) (accessed on 14 May 2026).

<sup>7</sup> Spandana Singh, "Everything in moderation", 22 July 2019, *New America*, <http://newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/>



- The technology's ability to manipulate user intentions and thought processes poses risks to the right to freedom of thought. The larger models and more expansive training data associated with generative AI models has led to the widening false perception that larger AI systems are equivalent to greater accuracy. However, users of more complex and more widely available AI systems, such as chat-based generative AI tools, are at particular risk of accepting generated outputs as reliable, given the increasing scale and centrality of generative AI products to digital querying, paving the way for an inhibition of critical faculties, or at worst, deliberate manipulation. Research in cognitive science, for instance, has found that frequency of exposure to fabricated information "predicts how deeply ingrained the belief in that information becomes".<sup>8</sup> Similarly, studies find that repeated exposure to algorithmic biases has a similar effect.
- These findings reflect design choices in the data pipeline that companies have made to prioritise scale and expedience, over critical human rights protections, which would have implications for the end product.

In Amnesty International's analysis, we find that machine learning methods used by major AI companies to develop standalone generative AI tools are based on a certain set of design choices in their training data input, processing, and scaling, which render them either in violation of particular principles of IHRL or put them at heightened risk of violating them. ***In our analysis, these design choices are not inevitable to generative AI*** but are adopted by companies that have chosen to rely on training data based on non-consensual and massive collection of publicly available data, including personal data, from the internet. The reliance on unlawful web scraping introduces numerous human rights challenges, risks and harms and is thus a design choice that must be challenged.

While LLM-based generative AI tools have been particularly popular since the launch of ChatGPT in November 2022, there are many other types and scales of generative AI technologies and techniques, many of which do not rely on unlawful web-scraping or mass invasions of privacy, such as, for instance, small language models (SLMs). These are often smaller in scope and size in terms of both training data and parameters (number of predictive models that decide what to do with the input). This means that SLM-based generative AI systems are often highly domain-specific and specialized, whereas LLMs can often veer into misleadingly appearing to have wide-ranging expertise. SLMs are consequently less resource intensive, more accurate, based on often local and industry-specific training data, and less prone to the problems associated with LLMs.<sup>9</sup>

This briefing focuses most extensively on standalone generative AI products that have been made publicly available and that rely on mass data collection practices by design. This includes large-scale web scraping to support the collection of training data, the production of synthetic data, and the generation of "new" outputs, the provenance of which is often left obscure and opaque by developers. This briefing does not cover "narrow" AI systems, such as facial recognition, predictive policing, and automated decision-making systems, for which Amnesty International has already documented significant human rights concerns in other publications.<sup>10</sup>

For the purposes of this briefing, we look at the models powering some of the most popular publicly available standalone generative AI tools, including ChatGPT, Dall-E, Gemini, Midjourney, LLaMa, Stable Diffusion, and DeepSeek. The models under analysis include large language models (LLMs), in particular generative pre-trained transformers (GPT), generative adversarial networks (GANs), variational autoencoders (VAEs) and diffusion models (DM).<sup>11</sup> (See Chapter 3)

The briefing concludes that standalone generative AI systems, based on unlawful web scraping, depend on mass invasions of privacy by design, and are fundamentally incompatible with IHRL. As such, Amnesty International is calling for a prohibition of such systems, including where such systems are identified as exacerbating existing inequalities or creating new forms of discrimination. Amnesty International's analysis

<sup>8</sup> Celeste Kidd and Abeba Birhane, "How AI can distort human beliefs", 22 June 2023, Science, Volume 380, Issue 6651, <https://doi.org/10.1126/science.adi0248>

<sup>9</sup> Bender, E.M. and others (2021) "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '21), pp. 610–623. Available at: <https://doi.org/10.1145/3442188.3445922>; Belcak, P. and others (2025) "Small Language Models are the Future of Agentic AI". arXiv. Available at: <https://doi.org/10.48550/arXiv.2506.02153>

<sup>10</sup> Amnesty International, "Ban the Scan", <http://banthescan.amnesty.org>; Amnesty International, 12 November 2024, "AI-powered welfare system fuels mass surveillance and risks discriminating against marginalized groups – report", <https://www.amnesty.org/en/latest/news/2024/11/denmark-ai-powered-welfare-system-fuels-mass-surveillance-and-risks-discriminating-against-marginalized-groups-report/>; Amnesty International UK, February 2025, "Automated Racism", <https://www.amnesty.org.uk/files/2025-02/Automated%20Racism%20Report%20-%20Amnesty%20International%20UK%20-%202025.pdf>

<sup>11</sup> "Standalone" designates generative AI systems that are available as distinct products, as opposed to generative AI features available in larger suites of products.

and conclusions demonstrate the urgent need to address the most harmful practices among AI companies, where disregard for international human rights standards has enabled the sector to spin-out high-risk products at the expense of personal data, equality and non-discrimination, labour conditions, environmental considerations, and cognitive health. Sophisticated data-intensive technologies, often labelled under “AI”, do not need to be built on harmful practices such as unlawful web scraping.

This briefing is intended as an advocacy and legal analysis tool to support civil society efforts to challenge the human rights abuses linked to generative AI companies and state deployment of these technologies. It seeks to support advocacy that confronts the concentration of power within AI companies, much of which stems out of unchecked data pipelines, with implications for other legal and regulatory areas such as intellectual property, anti-trust and market power, and non-consensual use of children’s data.

This is particularly urgent at a time when AI infrastructure is expanding rapidly across all parts of the supply chain. In this context, Amnesty International sees it as critical to bring a robust human rights analysis to this expansion, in order to prevent extractive practices from becoming entrenched in the development and deployment of AI systems.

# 2. TECHNICAL BACKGROUND TO GENERATIVE AI

In general, AI systems rely on complex supply chains that are each comprised of a network of actors responsible for the various aspects of the system's training and development.<sup>12</sup> These different actors are often tied together by data flows and together they produce the technology's functionality.<sup>13</sup>

For instance, a computing hardware manufacturer may provide server units to a cloud provider to set up data centres for the provision of cloud services. The cloud service provider may provide server space that would be "leased" to a client AI company intending to design and deploy an AI product; the cloud service provider may even go as far as providing Application Programming Interfaces (APIs)<sup>14</sup> that effectively allow the client company to build aspects of their AI product's functionality based on pre-compiled code.

AI systems are a product of labour, data, software and financial inputs with many layers of procurement embedded within the production process. This is not a novel problem, and companies have grappled with upstream and downstream harms of supply chains across industries for decades. However, in the case of AI, an extensive part of these supply chains is data and software driven – the *data pipeline* – rather than comprised of physical products, and often not linear in flow but rather through a web of actors and data movements. This opens opportunities for regulatory arbitrage, enables cross-border activities to happen in real time and for supply chains to grow to unprecedented scales.<sup>15</sup>

Generative AI systems are a type of AI that use machine learning models to produce new data or content by mimicking the input on which they have been trained. Generative AI is a blanket term used to describe a range of "deep learning" algorithmic models that are trained on vast amounts of data, and which then use this training data to generate new content, including audio, images, text, and even computer code in response to user prompts. In this briefing, the data collection (for training data), data processing (for machine learning), and data output aspects of generative AI systems are examined against international human rights standards. A brief description of some of the most common models discussed is outlined below.

---

<sup>12</sup> Ian Brown, Allocating accountability in AI supply chains, 29 June 2023, <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/> (accessed on 14 May 2026).

<sup>13</sup> Jennifer Cobbe, Michael Veale, and Jatinder Singh, Understanding accountability in algorithmic supply chains, 12 June 2023, <https://dl.acm.org/doi/10.1145/3593013.3594073> (accessed on 14 May 2026).

<sup>14</sup> A set of protocols that allow different applications, specifically software, to communicate and operate through one another. For example, 3<sup>rd</sup> party software that may be downloaded to or run through an existing piece of software.

<sup>15</sup> Jennifer Cobbe, Michael Veale and Jatinder Singh, 12 June 2023 (previously cited).

## 2.1 LARGE LANGUAGE MODELS (LLMs)

LLMs are machine learning models that rely on a vast corpus of content to produce, process and machine-interpret human inputs, often to produce other outputs. They are a form of “foundation model” (also referred to as “general purpose AI”); broad AI models capable of carrying out a vast number of diverse functions beyond conventional, narrow AI systems.<sup>16</sup> For example, an LLM might be used to have a conversation (such as a chatbot) but could also be used in the same instance to carry out large calculations, interpret and re-organize databases, and/or generate and process artwork and other imagery. Some of the most popular LLM-based generative AI products are based on generative pre-trained transformer models (GPTs), which are types of models that process and predict natural language (based on input “prompts”) for querying purposes.

While the most popular generative AI systems are powered by LLMs, small language models (SLMs) also power generative AI systems. These are often smaller in scope and size in terms of both training data and parameters (number of predictive models that decide what to do with the input). This means that SLM-based generative AI systems are often highly domain-specific and specialized, whereas LLMs can often veer into misleadingly appearing to have wide-ranging expertise. SLMs are consequently less resource intensive, more accurate, based on often local and industry-specific training data, and less prone to the problems associated with LLMs.<sup>17</sup>

While LLM-based generative AI tools have been particularly popular since the launch of ChatGPT in November 2022, there are many other types of generative AI technologies and techniques.

Most current generative AI systems are highly susceptible to unpredictability, inaccuracy and significant human rights risk.

## 2.2 GENERATIVE ADVERSARIAL MODELS (GANs)

A GAN is a model in which two neural networks – a generator and a discriminator – compete to become mutually more accurate in their predictions. The generator is tasked with generating synthetic data that could be mistaken for “real” data, while the discriminator is tasked with identifying which of the outputs have been artificially generated.

A feedback loop between the generator and the discriminator, in theory, develops more accurate outputs over time. Generator neural networks are often trained on massive training datasets of images to learn to generate synthetic versions of the input content. GANs are frequently used for real-time facial re-enactment (such as “deep fakes”) and to purportedly generate new configurations of “unique” faces.<sup>18</sup>

DALL-E, the public generative art tool released by OpenAI, is one of many products using GANs.

## 2.3 VARIATIONAL AUTOENCODERS (VAEs)

In simple terms, VAEs are generative AI models used to create “new” outputs by generating variations of the input data.<sup>19</sup> The model consists of:

1. An encoder that learns to identify and compress important features (called latent variables) from training data. For example, the encoder neural network might learn to recognize the key elements that make up a face.
2. A decoder that takes these learned features and either reconstructs the original input or creates new variations.<sup>20</sup>

---

<sup>16</sup> Eliot Jones, “What is a foundation model?” (previously cited).

<sup>17</sup> Bender, E.M. and others (2021) “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜”, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '21), pp. 610–623. Available at: <https://doi.org/10.1145/3442188.3445922>; Belcak, P. and others (2025) ‘Small Language Models are the Future of Agentic AI’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2506.02153>

<sup>18</sup> Mitra Azar and others, “Introduction: ways of machine seeing”, 20 February 2021, *AI & Society*, Volume 36, Issue 4, <https://doi.org/10.1007/s00146-020-01124-6>

<sup>19</sup> Diederik P. Kingma and Max Welling, “Auto-Encoding Variational Bayes”, 10 December 2022, arXiv, <https://doi.org/10.48550/arXiv.1312.6114>

<sup>20</sup> Dave Bergmann and Cole Stryker, “What is a variational autoencoder?”, IBM, <https://www.ibm.com/think/topics/variational-autoencoder> (accessed on 12 March 2025).

VAEs can generate smooth variations between different examples. For example, when trained on faces, VAEs can create images of new faces that blend features from multiple people.

Unlike traditional autoencoders, VAEs use probability distributions to represent their collection of latent variables (called “latent space”), which purportedly helps them to generate more realistic and diverse outputs. VAEs are frequently used in image generation, as well as for surveillance tools, such as anomaly detection (detecting movement/behaviour that appears to be outside of the “norm” for what an algorithm “expects”).

## 2.4 DIFFUSION MODELS

Diffusion models work by learning how to create new data by first understanding how data can be broken down.<sup>21</sup> They do this through a two-step process: first, “destroying” training data by gradually adding random data (also known as Gaussian noise) and then, learning to reverse this process to recover the original data.

The concept draws inspiration from physics, particularly the way particles move in nature. Imagine dropping ink into a glass of water – the way the ink particles gradually spread out (or diffuse) through the water follows a predictable pattern.<sup>22</sup> In diffusion models, pixels in images are treated like these ink molecules. By understanding this spreading-out process, the model learns how to reverse it, allowing it to generate new images from what initially looks like random noise. Diffusion models are also used in popular AI image generation tools like DALL-E and Stable Diffusion.

## 2.5 STANDALONE GENERATIVE AI TOOLS

Amnesty International defines standalone generative AI tools as products that are developed, deployed and marketed for their generative AI capabilities solely and specifically, such as AI chatbots, image/video/audio/text generators, and so on. This does not include products where generative AI is an added feature or function in a larger suite of products, for example word processing software with optional generative AI features. Standalone generative AI products are, in other words, generative AI models that come with their own front-end for direct use that is fundamentally concerned with generating outputs.

---

<sup>21</sup> Dave Bergmann and Cole Stryker, “What are diffusion models?”, IBM, <https://www.ibm.com/think/topics/diffusion-models> (accessed on 12 March 2025).

<sup>22</sup> Jonathan Ho and others, “Denoising diffusion probabilistic models”, 16 December 2020, arXiv, <https://doi.org/10.48550/arXiv.2006.11239>



# 3. RELEVANT LEGAL STANDARDS

This briefing draws on the following legal standards under IHRL to the design of prominent generative AI systems with implications for the rights under study in the briefing. Chiefly, it draws on the International Covenant on Civil and Political Rights (ICCPR), the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), alongside inputs from the UN special rapporteurs on contemporary forms of racism, privacy, and freedom of expression. It will also include reflections on business and human rights, including with reference to the UN Guiding Principles on Business and Human Rights (UN Guiding Principles), the OECD Guidelines for Multinational Enterprises (OECD Guidelines), and the UNDP Guide on Heightened Human Rights Due Diligence for Business in Conflict-Affected Contexts (UNDP Guide).

## 3.1 THE RIGHT TO PRIVACY

International human rights instruments, particularly Article 17 of the ICCPR, safeguard the right to privacy by prohibiting “arbitrary or unlawful interference” with one’s privacy, family, home or correspondence, requiring legal protection of these rights.<sup>23</sup>

Furthermore, children’s right to privacy is expressly protected under Article 16 of the Convention on the Rights of the Child (CRC). The UN Human Rights Committee (HRC), the body responsible for monitoring the implementation of the ICCPR by states parties, has long recognized that such protection includes regulating “the gathering and holding of personal information on computers, data banks and other devices, whether by public authorities or private individuals or bodies.”<sup>24</sup>

The UN Human Rights Committee (HRC) has consistently held that privacy protections extend to regulating “the gathering and holding of personal information on computers, data banks and other devices, whether by public authorities or private individuals or bodies”.<sup>25</sup>

The HRC has established that information and data available in “public areas” can be protected by Article 17.<sup>26</sup> As the UN High Commissioner for Human Rights (OHCHR) has further clarified: “the right to privacy comes into play when a government is monitoring a public space, such as a marketplace or a train station, thereby observing individuals... The public sharing of information does not render its substance unprotected.”<sup>27</sup>

The scope of privacy has always evolved in response to societal change, particularly new technological developments. The OHCHR has stated: “Privacy can be considered as the presumption that individuals should have an area of autonomous development, interaction and liberty, a ‘private sphere’ with or without

---

<sup>23</sup> UDHR, Article 12; ICCPR, Article 17.

<sup>24</sup> HRC, General Comment 16: The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation (Article 17), UN Doc. HRI/GEN/1/Rev.9 (Vol. I), 8 April 1988, para. 10.

<sup>25</sup> UN Human Rights Committee (HRC), General Comment 16: Article 17 (Right to Privacy), The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation, 8 April 1988, <https://www.refworld.org/docid/453883f922.html>

<sup>26</sup> HRC, Concluding Observations: Colombia, 17 November 2016, UN Doc. CCPR/C/COL/7, para. 32

<sup>27</sup> OHCHR, *The Right to Privacy in the Digital Age*, 3 August 2018, UN Doc. A/HRC/39/29, para. 6.

interaction with others, free from State intervention and from excessive unsolicited intervention by other uninvited individuals.”<sup>28</sup>

This encompasses three interrelated concepts: the freedom from intrusion into our private lives, the right to control information about ourselves, and the right to a space in which we can freely express our identities.<sup>29</sup>

Under IHRL, interference with privacy rights is only permissible when it meets strict legal requirements. A three-part test determines the legitimacy of such interference. First, the interference must be legally prescribed with clear, precise laws that include adequate safeguards and judicial oversight. Second, there must be a legitimate aim, such as protecting national security or public order. Third, the interference must be both necessary and proportionate to achieve this aim. This means using the least restrictive measure possible and ensuring that the benefits outweigh any potential harm. Additionally, any discriminatory interference is automatically considered unlawful and arbitrary under international law.

Amnesty International has long held that AI-driven surveillance tools, such as facial recognition technology (FRT), that scans, captures and often stores data on massive databases curated without individuals’ knowledge and consent, are tools of mass surveillance by design. Amnesty International believes that indiscriminate mass surveillance, including through FRT, is never a proportionate interference with the right to privacy. Similarly, Amnesty International’s analysis concludes that the advertising-driven business model of large tech companies such as Google and Facebook creates an unprecedented interference with the right to privacy that cannot be compatible with the companies’ responsibility to respect human rights.<sup>30</sup>

Through similar processes, including the use of web scrapers for collecting training data, some generative AI tools risk violating privacy laws and the rights enshrined within them. This may depend on whether the collection and processing of data used to build the tool has been conducted consensually or whether the individual and community right to privacy has been violated in the process, whether it serves a legitimate interest, whether it is accurate and secure, and whether individuals may exercise their rights under data protection law. The right to privacy is therefore crucial in understanding the impact of generative AI technologies on human rights.

## 3.2 THE RIGHT TO EQUALITY AND NON-DISCRIMINATION

The right to equality and non-discrimination is a central principle that underpins all human rights.<sup>31</sup> It is protected through various international human rights instruments, notably the ICERD and the ICCPR. The UN Committee on the Elimination of Racial Discrimination (ICERD Committee) has emphasized that equality should be interpreted broadly, encompassing both formal equality before the law and substantive equality in the practical exercise and enjoyment of human rights.<sup>32</sup> The prohibition of racial discrimination under ICERD applies to both direct and indirect discrimination, the latter of which implicates measures that seem neutral on face value but lead to discriminatory results. The prohibition on racial discrimination applies to all states independent of their treaty obligations, and creates obligations that are owed to the international community as a whole that states must fulfil with no exceptions.<sup>33</sup> Racial discrimination is defined broadly in IHRL to include discrimination on the basis of race, colour, descent or national or ethnic origin.<sup>34</sup> Discrimination on the grounds of gender, sex and religion are, like racial discrimination, are also prohibited under IHRL.<sup>35</sup> The right to equality and non-discrimination also applies to the rights of children. The rights of children to be

<sup>28</sup> OHCHR, *The Right to Privacy in the Digital Age* (previously cited), para. 5.

<sup>29</sup> Amnesty International, *Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights* (Index: POL 30/1404/2019), 21 November 2019, <https://www.amnesty.org/en/documents/pol30/1404/2019/en/>

<sup>30</sup> Amnesty International, *Surveillance Giants* (previously cited), p. 22.

<sup>31</sup> “Non-discrimination and equality are fundamental components of international human rights law and essential to the exercise and enjoyment of economic, social and cultural rights.” – UN Committee on Economic, Social and Cultural Rights (CESCR), General Comment 20: Non-discrimination in economic, social and cultural rights, 2 July 2009, UN Doc. E/C.12/GC/20, para. 2; “Non-discrimination, together with equality before the law and equal protection of the law without any discrimination, constitute a basic and general principle relating to the protection of human rights.” – HRC, General Comment 18: Non-discrimination, 1989, UN Doc. HRI/GEN/1/Rev.9(Vol. I), p. 95, para. 1.

<sup>32</sup> Committee on the Elimination of Racial Discrimination (CERD Committee), General Recommendation 32, 24 September 2009, CERD/C/GC/32, para. 6.

<sup>33</sup> The prohibition on racial discrimination is a peremptory norm of customary international law (also known as *jus cogens*), which means that it applies to all states independently of their treaty obligations, and gives rise to obligations *erga omnes* (obligations that are owed to the international community as a whole) from which states cannot derogate. *Case Concerning the Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain)* (Judgment), International Court of Justice, Rep 3 (1970), paras 33–34. See also, International Law Commission (ILC), “Draft conclusions on identification and legal consequences of peremptory norms of general international law (*jus cogens*)”, *Yearbook of the International Law Commission*, 2022, Volume II, Part Two, Conclusion 23 & Annex.

<sup>34</sup> ICERD, Article 1(1); CERD/CERD Committee, General Recommendation 32 (previously cited), para. 6.

<sup>35</sup> See for example, ICCPR, CEDAW and the Declaration on the Elimination of All Forms of Intolerance and Discrimination Based on Religion or Belief; HRC, General Comment 18 (previously cited), para. 7; Daniel Moeckli and others (eds), *International Human Rights Law*, 3rd edition, 2018, p. 151; Malcolm Shaw, *International Law*, 8th edition, 2017, p. 287.

protected from all forms of discrimination is expressly reiterated in Article 2 of the CRC. Furthermore, the Committee on the Rights of the Child has identified the obligation of states to ensure that all children within their jurisdiction enjoy the rights laid out in the CRC without discrimination of any kind as one of the four general principles in light of which all the rights enshrined in the CRC should be interpreted and implemented.<sup>36</sup>

According to the former UN Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, E. Tendayi Achiume, digital technologies have been used in ways that “produce racially discriminatory structures that holistically or systematically undermine enjoyment of human rights for certain groups, on account of their race, ethnicity or national origin, in combination with other characteristics.”<sup>37</sup> In light of this, the Special Rapporteur contended that outright bans on certain digital technologies may be required to prevent racially discriminatory outcomes and other human rights violations until such risks can be mitigated.<sup>38</sup>

In 2024, the current UN Special Rapporteur on contemporary forms of racism, Ashwini K.P.,<sup>39</sup> highlighted the compounding discriminatory nature of generative AI technologies in particular. The Special Rapporteur emphasised biased outcomes on the basis of over- or under-representation along lines of race and ethnicity in training data, and highlighted the reliance on systems built on incomplete or synthetic data, and their reproduction of racial discrimination.<sup>40</sup> The Special Rapporteur reiterates the call on states to “consider prohibiting the use of artificial intelligence systems that have been shown to have unacceptable human rights risks, including those that violate the prohibition of racial discrimination,” while calling on private actors to “undertake human rights due diligence assessments at all stages of artificial intelligence product design, development and deployment” and to “develop protocols for ensuring full transparency and the sharing of information about algorithmic decision-making for products that have human rights implications.”<sup>41</sup>

The right to equality and non-discrimination also prohibits discrimination on the basis of gender identity and sex characteristics, and sexual orientation.<sup>42</sup> Article 1 of the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) defines discrimination as: “any distinction, exclusion or restriction made on the basis of sex which has the effect or purpose of impairing or nullifying the recognition, enjoyment or exercise by women, irrespective of their marital status, on the basis of equality of men and women, of human rights and fundamental freedoms in the political, economic, social, cultural, civil or any other field.”

In 2023 the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan, explored gendered disinformation as a strategy for targeting women and gender-nonconforming individuals. Of particular relevance is the argument that “[u]nlike other forms of disinformation, gendered disinformation relies not just on false information *but also on existing gender narratives* to achieve its social and political goals, including maintaining the status quo of gender or creating a more polarized electorate”.<sup>43</sup> This also applies to systems that would seek to, in direct or indirect ways, reinforce and replicate such dynamics of discrimination.

In the context of children, Article 34 of the UN Convention on the Rights of the Child (CRC)<sup>44</sup> is the main provision on protection of children from sexual exploitation and abuse, which applies to child sexual abuse material (CSAM), including AI-generated CSAM. It states:

**States Parties undertake to protect the child from all forms of sexual exploitation and sexual abuse. For these purposes, States Parties shall in particular take all appropriate national, bilateral and multilateral measures to prevent:**

---

<sup>36</sup> UN Committee on the Rights of the Child, General Comment 5: General Issues of Implementation of the Convention on the Rights of the Child (Articles 4, 42 and 44, para. 6), 27 November 2003, UN Doc. CRC/GC/2003/5 and General Comment 14: The Right of the Child to Have His or Her Best Interests Taken as a Primary Consideration (Article 3, para. 1), 29 May 2013, UN Doc. CRC /C/GC/14.

<sup>37</sup> Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance: *Racial discrimination and emerging digital technologies: a human rights analysis*, 18 June 2020, UN Doc. A/HRC/44/57, paras 38-43.

<sup>38</sup> Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, 2020, (previously cited), para. 56.

<sup>39</sup> OHCHR, Ms. Ashwini K.P., Special Rapporteur on contemporary forms of racism, <https://www.ohchr.org/en/special-procedures/sr-racism/ms-ashwini-kp> (accessed on 21 May 2026)

<sup>40</sup> Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance: *Contemporary forms of racism, racial discrimination, xenophobia and related intolerance*, 3 June 2024, UN Doc. A/HRC/56/68.

<sup>41</sup> Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, 2024, (previously cited).

<sup>42</sup> See ICCPR, Article 2; HRC, *Toonen v. Australia*, Decision, 31 March 1994, <https://juris.ohchr.org/casedetails/702/en-US>

<sup>43</sup> Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression: *Gendered disinformation and its implications for the right to freedom of expression*, 7 August 2023, UN Doc. A/78/288.

<sup>44</sup> Adopted on 20 November 1989, entered into force on 2 September 1990.

- a) **The inducement or coercion of a child to engage in any unlawful sexual activity;**
- b) **The exploitative use of children in prostitution or other unlawful sexual practices;**
- c) **The exploitative use of children in pornographic performances and material.**

Virtual CSAM, including AIG-CSAM, is not explicitly criminalised under the Conventions on the Rights of the Child and Optional Protocol on the Sale of Children, Child Prostitution, and Child Pornography (OPSC), even though its criminalisation has been recommended by the CRC Committee since 2019.

The CRC Committee in its non-binding 2019 Guidelines on the OPSC expressed its deep concern about the “large amount of online and offline material, including drawings and virtual representations, depicting non-existing children or persons appearing to be children involved in sexually explicit conduct, and about the serious effect that such material can have on children’s right to dignity and protection.” Therefore, the Committee encourages States parties to include “non-existing children or persons appearing to be children” under child sexual abuse material provisions.<sup>45</sup>

Just as AI-driven surveillance tools such as facial recognition have been subject to calls for prohibition due to the high risk of profiling according to ethnicity, race, national origin, gender identity and other prohibited grounds, which is often the basis for unlawful discrimination, generative AI tools that rely on massive data collection exercises present similar risks of discrimination.<sup>46</sup>

### 3.3 THE RIGHT TO FREEDOM OF EXPRESSION

States have a duty to safeguard the right to seek, receive and impart information and ideas of all kinds – including political, religious or philosophical, artistic and cultural information and ideas – by any means, regardless of frontiers.<sup>47</sup> The right to privacy is “an essential requirement for the realization of the right to freedom of expression.”<sup>48</sup> Freedom of expression is a collective right, enabling people to seek and receive information as a social group and to “voice their collective views.”<sup>49</sup>

The right to free expression, however, is not absolute. Article 19.3 of the ICCPR states that, for a restriction of the right to be permissible, it must be provided by law and be necessary (including proportionate) to meet one of the enumerated legitimate aims (protecting the rights or reputations of others, national security or public order, or public health or morals). Protecting the rights of others includes, among others, the right to be protected against incitement to discrimination.<sup>50</sup>

Accordingly, under Article 20 of the ICCPR, states are required to prohibit (though not necessarily criminalize) “propaganda for war” and “advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.” Authorities may only criminalize advocacy of hatred if it reaches a high threshold based on a cumulative test which requires (among other factors) an assessment of the context; the position or status of the speaker in society; the intent of the speaker; the size of the audience; and the likelihood that there is a reasonable probability that the speech would succeed in inciting actual action against the target group.

Private actors also have a responsibility to respect human rights, as established under international human rights standards such as the UN Guiding Principles. Content moderation on the internet, including social media platforms, can be considered one such measure by which private actors can fulfil their responsibility to respect human rights in the context of free expression. Content moderation must be conducted in a manner ensuring respect for the right to freedom of expression, while addressing rampant advocacy of hatred – a systemic issue on social media platforms.

<sup>45</sup> CRC Committee, *Guidelines regarding the implementation of the Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography*, CRC/C/156, para. 63.

<sup>46</sup> Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression: *Surveillance and Human Rights*, 28 May 2019, UN Doc. A/HRC/41/35, para. 12.

<sup>47</sup> UDHR, Article 19; ICCPR, Article 19.

<sup>48</sup> Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, 17 April 2013, UN Doc. A/HRC/23/40, para. 24.

<sup>49</sup> Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, 20 April 2010, UN Doc. A/HRC/14/23, para. 29.

<sup>50</sup> UDHR, Article 7.

## 3.4 FREEDOM OF THOUGHT

The right to freedom of thought is enshrined in Article 18 of the Universal Declaration of Human Rights (UDHR) and Article 18 of the ICCPR, which stipulates “that everyone shall have the right to freedom of thought, conscience and religion.” Of particular relevance is Article 18(2) of the ICCPR, which ensures protection from coercion that could impair an individual’s freedom to “have or to adopt a religion or belief of his choice”. Freedom of thought extends more specifically to the “right to keep our thoughts private; the right to keep our thoughts free from manipulation; and the right not to be penalized for our thoughts alone”.<sup>51</sup>

In 2021, the then UN Special Rapporteur on freedom of religion or belief, Ahmed Shaheed, presented his *Report on Freedom of Thought*, in which he drew particular attention to emerging scholarship that warned against developments in “digital technology, neuroscience, and cognitive psychology that could potentially enable access to the very content of our thoughts and affect how we think, feel and behave.”<sup>52</sup>

International legal opinion on the right to freedom of thought has long held that the right is considered absolute and cannot be legally interfered with, “even during public emergencies”.<sup>53</sup> In his 2021 report, Special Rapporteur Ahmed Shaheed describes the right as being “unlike forum externum (external realm) freedoms that are subject to State limitations, if prescribed by law and necessary to protect public safety, order, health or morals, or the rights of others, States legally cannot ever interfere with freedom of thought”.<sup>54</sup>

Importantly, the Special Rapporteur highlights legal scholarship that delineates what possible violations of Article 18 could look like. The report highlights *coercion* (psychological influence that could “include coercive alterations of thought,” including through use of force or implied threat), *modification* (changing of thought via “direct alteration of brain chemistry or brain function... irrespective of the victim’s awareness of the use or threat of force”), and *manipulation* (interference with processes of understanding, which could generate biased mental models, knowledge and ideology), as actions through which violations of the right to freedom of thought could be carried out. In particular, on actions related to manipulation, the extent to which power differentials are exploited by a malicious actor plays a significant role in establishing to what extent a violation has occurred.

In a 2024 article, legal scholars Patrick O’Callaghan and others highlight the serious questions raised for freedom of thought by “the impact of intrusive data collection and surveillance activities”.<sup>55</sup> They argue that the EU’s Digital Services Act (DSA) and Artificial Intelligence Act (AI Act) carry the regulatory means for realizing the right to freedom of thought as a function of technology-driven violations.

Article 25(1) of the DSA prohibits the deceptive and manipulative practice of “Dark Patterns” (deceptive user interface designs that manipulate users into making certain choices or taking certain actions),<sup>56</sup> and holds that tech actors should not “materially [distort] or [impair] the ability of the recipients of their service to make free and informed decisions”.<sup>57</sup> Article 26(3) furthermore prohibits the use of targeted advertising based on user profiling that invokes special categories of personal data as inscribed in Europe’s General Data Protection Regulation (GDPR). Finally, Article 34 requires ongoing assessment of systemic risks stemming from “the design of operation of their service and systems, including their recommender and advertising systems.” Patrick O’Callaghan and others note that, because this stipulation is specifically concerned with “any actual or foreseeable negative effects for the exercise of *fundamental rights*”, it should extend to the right to freedom of thought upheld in Article 10, as well as Article 3 (the right to mental integrity), of the Charter of Fundamental Rights of the European Union.<sup>58</sup>

Article 5 of the AI Act furthermore stipulates the following prohibitions on AI practices that relate to the right to freedom of thought:

---

<sup>51</sup> Susie Alegre and Aaron Shull, *Freedom of Thought: Reviving and Protecting a Forgotten Human Right*, 2024, Centre for International Governance Innovation, [https://www.cigionline.org/static/documents/Freedom.of.Thought\\_SpecialReport.Alegre.Shull.pdf](https://www.cigionline.org/static/documents/Freedom.of.Thought_SpecialReport.Alegre.Shull.pdf)

<sup>52</sup> Interim report of the Special Rapporteur on freedom of religion or belief, Ahmed Shaheed: *Freedom of Thought*, 5 October 2021, UN Doc. A/76/380.

<sup>53</sup> HRC, General Comment 22, on Article 18, 27 September 1993, UN Doc. CCPR/C/21/Rev.1/Add.4, paras 1 and 3. See also HRC, General Comment 34, Article 19: Freedoms of opinion and expression, 12 September 2011, UN Doc. CCPR/C/GC/34, para.5; Report of the Special Rapporteur on freedom of religion or belief, 23 December 2015, UN Doc. A/HRC/31/18, <https://docs.un.org/en/A/HRC/31/18> para. 17.

<sup>54</sup> Interim report of the Special Rapporteur on freedom of religion or belief, Ahmed Shaheed: *Freedom of Thought* (previously cited).

<sup>55</sup> Patrick O’Callaghan and others, “The right to freedom of thought: an interdisciplinary analysis of the UN special rapporteur’s report on freedom of thought” 2024, *The International Journal of Human Rights*, Volume 28, Issue 1, <https://doi.org/10.1080/13642987.2023.2227100>

<sup>56</sup> Patrick O’Callaghan and others, “The right to freedom of thought: an interdisciplinary analysis of the UN special rapporteur’s report on freedom of thought” (previously cited).

<sup>57</sup> Patrick O’Callaghan and others, “The right to freedom of thought: an interdisciplinary analysis of the UN special rapporteur’s report on freedom of thought” (previously cited).

<sup>58</sup> Patrick O’Callaghan and others, “The right to freedom of thought: an interdisciplinary analysis of the UN special rapporteur’s report on freedom of thought” (previously cited).



- A. “the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person’s consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm”;
- B. “the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm”.

A 2024 report published by the Centre for International Governance Innovation draws attention to the right to freedom of thought as an increasingly salient right to be considered against the backdrop not only of emerging brain-computer interfaces, but also data-driven persuasive technologies, eye-tracking and pupillometry (measuring the movement and reaction of the eye), scene understanding (ability to comprehend and analyse a visual scene), and the predictive power of AI language models.<sup>59</sup> While the right to freedom of thought is an oft-neglected right in international law, it has found its way to hard law applications, and is an increasingly salient and important tool in assessing the rapid adoption of generative AI technologies in particular.

## 3.5 BUSINESS AND HUMAN RIGHTS STANDARDS

Companies have a responsibility to respect all human rights wherever in the world they operate and throughout their operations. This is a widely recognized standard of expected conduct as set out in international business and human rights standards, including the UN Guiding Principles and the OECD Guidelines for Multinational Enterprises (OECD Guidelines).<sup>60</sup> The UN Guiding Principles also make clear that companies have a responsibility to respect standards of international humanitarian law.<sup>61</sup> The OHCHR has explained that international humanitarian law imposes obligations on business managers and staff not to breach the rules of international humanitarian law.<sup>62</sup>

The corporate responsibility to respect human rights and international humanitarian law is independent of a state’s own human rights obligations and exists over and above compliance with national laws and regulations protecting human rights.<sup>63</sup> UN Guiding Principle 13 outlines that companies’ responsibility to respect human rights entails a requirement to “avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur; and seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts”.

To meet their corporate responsibility to respect human rights, companies should have in place ongoing and proactive human rights due diligence processes to identify, prevent, mitigate and account for how they address their impacts on human rights. Due diligence is based on the concept of proportionality: the more severe the risk, the more adapted the due diligence processes must be to the context and particularity of the risk.<sup>64</sup> Severity of impacts is assessed by their scale, scope and irremediable character.<sup>65</sup>

When conducting human rights due diligence, a company may identify that it may cause or contribute to – or already be causing or contributing to – a serious human rights abuse through its own activities. In these

<sup>59</sup> Susie Alegre and Aaron Shull, *Freedom of Thought: Reviving and Protecting a Forgotten Human Right* (previously cited).

<sup>60</sup> This responsibility was expressly recognized by the UN Human Rights Council on 16 June 2011, when it endorsed the UN Guiding Principles on Business and Human Rights (UN Guiding Principles), and on 25 May 2011, when the 42 governments that had then adhered to the Declaration on International Investment and Multinational Enterprises of the OECD unanimously endorsed a revised version of the OECD Guidelines for Multinational Enterprises. See Human Rights Council, Resolution 17/4: Human Rights and Transnational Corporations and other Business Enterprises, 6 July 2011, UN Doc. A/HRC/RES/17/4; OECD Guidelines for Multinational Enterprises, 2011, [www.oecd.org/corporate/mne](http://www.oecd.org/corporate/mne)

<sup>61</sup> UN Guiding Principles, Principle 12 including Commentary.

<sup>62</sup> OHCHR, *The Corporate Responsibility to Respect Human Rights: An Interpretive Guide*, 2012, [https://www.ohchr.org/sites/default/files/Documents/publications/hr.pub.12.2\\_en.pdf](https://www.ohchr.org/sites/default/files/Documents/publications/hr.pub.12.2_en.pdf)

See also, ICRC, *Business and International Humanitarian Law: An Introduction to the Rights and Obligations of Business Enterprises under International Humanitarian Law*, 2006, <https://www.icrc.org/en/publication/0882-business-and-international-humanitarian-law-introduction-rights-and-obligations>

<sup>63</sup> UN Guiding Principles, Principle 11 including Commentary.

<sup>64</sup> UNDP, *Heightened Human Rights Due Diligence for Business in Conflict-Affected Contexts: A Guide*, 16 June 2022, <https://www.undp.org/publications/heightened-human-rights-due-diligence-business-conflict-affected-contexts-guide>

<sup>65</sup> UN Guiding Principles, Principle 14 including Commentary.

cases, companies must cease or prevent the activities that are responsible for those adverse human rights impacts. Where abuses are outside of the business enterprise's control but are directly linked to their operations, products or services through their business relationships, the UN Guiding Principles require the company to seek to mitigate the human rights impact by exercising leverage, or by seeking to improve leverage where leverage is limited, including through collaboration if appropriate. If companies are not able to exercise leverage – or the leverage exercised is insufficient to mitigate the harm – then a company must responsibly disengage from the business relationship.

Transparency is a key component of human rights due diligence. As the UN Guiding Principles make clear, companies “need to know and show that they respect human rights”,<sup>66</sup> where “showing involves communication, providing a measure of transparency and accountability to individuals or groups who may be impacted and to other relevant stakeholders”.<sup>67</sup>

The Committee on the Rights of the Child has also set out measures (in addition to those on due diligence above) required of states to protect children's human rights in the context of corporate activities. These measures include adopting laws, regulations and policies to protect children's rights and best interests and monitoring their enforcement as well as investigating, adjudicating and ensuring redress for abuses of children's rights caused or contributed to by a business enterprise. States are considered “responsible for infringements of children's rights caused or contributed to by business enterprises where it has failed to undertake necessary, appropriate, and reasonable measures to prevent and remedy such infringements or otherwise collaborated with or tolerated the infringements.”<sup>68</sup>

---

<sup>66</sup> UN Guiding Principles, Commentary to Principle 15.

<sup>67</sup> UN Guiding Principles, Commentary to Principle 21.

<sup>68</sup> Committee on the Rights of the Child, General Comment 16, 17 April 2013, UN Doc. CRC/C/GC/16, para. 28.

# 4. BACKGROUND TO GLOBAL DEVELOPMENTS IN GENERATIVE AI

On 30 November 2022, the US-based company OpenAI released an early demo of ChatGPT; an AI-driven chatbot capable of processing and responding to users in natural speech.<sup>69</sup> While there is nothing new about AI tools or publicly available chatbots, publicly available generative AI models are capable of carrying out a vast number of diverse functions beyond conventional, “narrow” AI systems.<sup>70</sup> For example, while a generative AI chatbot might be used to simulate a conversation, the same tool could also be used to carry out calculations, interpret and reorganize databases, and generate and process artwork and imagery. Generative AI systems are, in other words, multi-modal in both their inputs and outputs, presenting a new form of processing that allows many different types of tasks to be carried out within the same system. This is possible due to the “foundational model” at the core of this new generation of AI, such as LLMs. LLMs were designed to process natural language and can be used across a great variety of tasks including content or code generation, summarization, mimicking conversation, creative writing and more. Since launching ChatGPT, OpenAI has integrated a number of other generative AI models into ChatGPT, including DALL-E,<sup>71</sup> a text-to-image generator, while also launching a hyper-realistic text-to-video generator named Sora.<sup>72</sup>

In addition to acceleration caused in part by the public launch of OpenAI’s GPT-based products, in the last five years, tech corporations have shifted their focus more fundamentally to AI development and deployment, with companies rapidly announcing generative AI tools, new cloud infrastructure offerings to support AI products, and initiatives that purport to keep AI harms in check. In November 2023, x.AI announced its Grok model,<sup>73</sup> closely followed by Google, which announced its Gemini model,<sup>74</sup> and Meta’s LLaMA model,<sup>75</sup> in December 2023. In January 2025, DeepSeek, an AI startup based in China, released an AI model named DeepSeek R1.<sup>76</sup>

This increasing speed of development is leading to widescale deployment of generative AI in all domains, all while AI regulation and safeguards lag behind and are largely unimplemented. In the highest stakes contexts, widescale adoption of generative AI has introduced a growing list of harmful outputs and outcomes, concerning from a rights-perspective. These include hostile and violent conversations between humans and chatbots,<sup>77</sup> fake and explicit AI-generated imagery of people – including children, young people and

---

<sup>69</sup> OpenAI, “Introducing ChatGPT”, 30 November 2022, <https://openai.com/index/chatgpt/>

<sup>70</sup> Eliot Jones, “What is a foundation model?”, 17 July 2023, Ada Lovelace Institute, <https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/>

<sup>71</sup> OpenAI, “DALL-E 2”, 2022, <https://openai.com/index/dall-e-2/>

<sup>72</sup> OpenAI, “Sora”, <https://openai.com/sora/> (accessed on 19 March 2025).

<sup>73</sup> *Announcing Grok / xAI* (November 2023). Available at: <https://x.ai/news/grok> (Accessed: 24 March 2026).

<sup>74</sup> Google, “Introducing Gemini: Google’s most capable AI model yet”, 6 December 2023, <https://blog.google/technology/ai/google-gemini-ai/>

<sup>75</sup> Meta, “Introducing LLaMA: A foundational, 65-billion-parameter language model”, 24 February 2023, <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>

<sup>76</sup> DeepSeek API Docs, “DeepSeek-R1 release”, 20 January 2025, <https://api-docs.deepseek.com/news/news250120>

<sup>77</sup> Bobby Allyn, “Microsoft’s new AI chatbot has been saying some ‘crazy and unhinged things’”, 2 March 2023, NPR, <https://www.npr.org/2023/03/02/1159895892/ai-microsoft-bing-chatbot>

celebrities<sup>78</sup> – harmful racial stereotypes and biased cultural representations,<sup>79</sup> maltreatment of data labellers involved in the AI supply chain,<sup>80</sup> exponential environmental cost associated with the increasing adoption of AI,<sup>81</sup> and surveillance and capture of massive amounts of personal, behavioural and creative data and content used to “train” AI models,<sup>82</sup> usually without the originators’ knowledge or consent.<sup>83</sup>

At its most extreme, there are also increasing reports of the potentially unlawful usage of generative AI tools in the military realm. In the occupied Palestinian territory (OPT), investigative reports revealed the devastating use of AI for target acquisition (detecting and identifying targets for military operations) in the context of the genocide in Gaza, through the Gospel system and later Lavender and Where’s Daddy.<sup>84</sup> In January 2024 the surveillance company Palantir also announced that it had been supplying Israeli authorities with new tools since the start of Israel’s genocide in Gaza in October 2023.<sup>85</sup> Palantir’s current military offering is its Artificial Intelligence Platform for Defense (AIP for Defense),<sup>86</sup> which operationalize LLMs in military contexts for decision-making.<sup>87</sup> Despite wide reporting on Israel’s genocide against Palestinians in Gaza,<sup>88</sup> including advisory opinions and provisional measures by the International Court of Justice, and a finding of genocide by the UN Commission of Inquiry on the Occupied Palestinian Territory, including East Jerusalem, and Israel, Israel’s military campaign continues to draw various AI investments to its cause.<sup>89</sup> Google and Amazon have come under heavy scrutiny and criticism, including from their own workers,<sup>90</sup> for providing the cloud computing platform Nimbus to the Israeli Ministry of Defence.<sup>91</sup> In January 2025, Google also dropped its commitment to not pursue the development of AI for weaponry, surveillance and other harms that contravene international law.<sup>92</sup> Similarly, Meta announced in November 2024 that it would allow US national security agencies and defence contractors to use its LLaMa model, in spite of previous principles published by the company outlining its prohibitions against such uses.<sup>93</sup> In February 2026, reports emerged that the US military had used Anthropic’s AI model, Claude, as part of its joint strikes

<sup>78</sup> Kate Conger and John Yoon, “Explicit deepfake images of Taylor Swift elude safeguards and swamp social media”, 26 January 2024, New York Times, <https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html>

<sup>79</sup> Ameera Kawash, “What a cat in a keffiyeh reveals about AI’s anti-Palestinian bias”, 25 April 2023, +972 Magazine, <https://www.972mag.com/ai-bias-palestinian-cat-keffiyeh/>

Liz O’Sullivan and John P. Dickerson, “Here are a few ways GPT-3 can go wrong”, 7 August 2020, TechCrunch, <https://techcrunch.com/2020/08/07/here-are-a-few-ways-gpt-3-can-go-wrong/>

Geoffrey Currie and others, “Gender and ethnicity bias in generative artificial intelligence text-to-image depiction of pharmacists”, December 2024, International Journal of Pharmacy Practice, Volume 32, Issue 6, <https://doi.org/10.1093/ijpp/riae049>

Ananya, “AI image generators often give racist and sexist results: can they be fixed?”, 19 March 2024, Nature, Volume 627, Issue 8005, <https://doi.org/10.1038/d41586-024-00674-9>

<sup>80</sup> Lesley Stahl and others, “Kenyan workers with AI jobs thought they had tickets to the future until the grim reality set in”, 24 November 2024, CBS News, <https://www.cbsnews.com/news/ai-work-kenya-exploitation-60-minutes/>

<sup>81</sup> Ana Valdivia, “The supply chain capitalism of AI: a call to (re)think algorithmic harms and resistance through environmental lens”, 30 Oct 2024, Information, Communication & Society, Volume 28, Issue 12, <https://doi.org/10.1080/1369118X.2024.2420021>

<sup>82</sup> Xiongbiao Ye and others, “Privacy and personal data risk governance for generative artificial intelligence: a Chinese perspective”, November 2024, Telecommunications Policy, Volume 48, Issue 10, <https://doi.org/10.1016/j.telpol.2024.102851>

Alice Nunwick, “OpenAI’s ‘web scraping’ still not compliant, warns EU regulator”, 24 May 2024, Verdict, <https://www.verdict.co.uk/openai-web-scraping-still-not-compliant-warns-eu-regulator/>

<sup>83</sup> Ernestas Naprys, “Meta leached 82 terabytes of pirated books to train its Llama AI, documents reveal”, 7 February 2025, Cybernews, <https://cybernews.com/tech/meta-leached-82-terabytes-of-pirated-books-to-train-its-llama-ai-documents-reveal/>

<sup>84</sup> Harry Davies and others, “The Gospel: how Israel uses AI to select bombing targets in Gaza”, 1 December 2023, The Guardian, <https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets>

Yuval Abraham, “Lavender: the AI machine directing Israel’s bombing spree in Gaza”, 3 April 2024, +972 Magazine, <https://www.972mag.com/lavender-ai-israeli-army-gaza/>

<sup>85</sup> Marissa Newman, “Palantir allegedly supplying Israel with AI tools amid Israel’s war in Gaza”, 10 January 2024, Business & Human Rights Resource Centre, <https://www.business-humanrights.org/en/latest-news/palantir-allegedly-supplying-israel-with-ai-tools-amid-israels-war-in-gaza/>

<sup>86</sup> Palantir, “Palantir AIP for Defense”, <https://www.palantir.com/platforms/aip/defense/> (accessed on 21 March 2025).

<sup>87</sup> Amnesty International, Pull the plug on the political economy enabling Israel’s crimes: What states and companies must do to stop fueling Israel’s genocide, apartheid and unlawful occupation, 18 September 2025, <https://www.amnesty.org/en/documents/pol40/0289/2025/en/>

<sup>88</sup> Amnesty International, “Amnesty International investigation concludes Israel is committing genocide against Palestinians in Gaza”, 5 December 2024, <https://www.amnesty.org/en/latest/news/2024/12/amnesty-international-concludes-israel-is-committing-genocide-against-palestinians-in-gaza/>

UN Special Rapporteur on the situation of human rights in the Palestinian territories occupied since 1967, Report, 25 March 2024, UN Doc. A/HRC/55/73; Forensic Architecture, “A cartography of genocide: Israel’s conduct in Gaza since October 2023”, <https://forensic-architecture.org/investigation/a-cartography-of-genocide> (accessed on 7 January 2026); Al-Haq, “Al-Haq publishes new report: ‘The Systematic Destruction of Gaza’s Healthcare System: A Pattern of Genocide’”, 23 January 2025, <https://www.alhaq.org/publications/25846.html>

<sup>89</sup> ICJ, *Legal Consequences arising from the Policies and Practices of Israel in the Occupied Palestinian Territory, including East Jerusalem*, Press release 2024/57, 19 July 2024, <https://www.icj-cij.org/case/186>

<sup>90</sup> Michael Sainato, “Workers accuse Google of ‘tantrum’ after 50 fired over Israel contract protest”, 27 April 2024, The Guardian, <https://www.theguardian.com/technology/2024/apr/27/google-project-nimbus-israel>

<sup>91</sup> Sam Biddle, “Israeli weapons firms required to buy cloud services from Google and Amazon”, 1 May 2024, The Intercept, <https://theintercept.com/2024/05/01/google-amazon-nimbus-israel-weapons-arms-gaza/>

<sup>92</sup> Zena Assaad, “Google has dropped its promise not to use AI for weapons. It’s part of a troubling trend”, 10 February 2025, The Conversation, <http://theconversation.com/google-has-dropped-its-promise-not-to-use-ai-for-weapons-its-part-of-a-troubling-trend-249169>

<sup>93</sup> Johana Bhuiyan, J. (2024) “Meta to let US national security agencies and defense contractors use Llama AI”, 5 November 2024, The Guardian, 5 November. Available at: <https://www.theguardian.com/technology/2024/nov/05/meta-allows-national-security-defense-contractors-use-llama-ai> (Accessed: 21 March 2025).

on Iran with Israel.<sup>94</sup> This came after the US government ignored contractual limitations instated by Anthropic on the use of its product for “mass domestic surveillance”, and “fully autonomous weapons [...] without oversight”.<sup>95</sup> Anthropic’s products have been used by the US government and military since 2024, and reportedly the first “advanced AI company” with products deployed across government agencies engaged in classified work.<sup>96</sup> Anthropic was labelled a “Supply Chain Risk” by the Pentagon on 6 March 2026, following their refusal to amend these limitations.<sup>97</sup>

States including India, Singapore, Taiwan, Denmark and the Netherlands are also increasingly pursuing “sovereign” AI projects through the pursuit of homegrown LLMs.<sup>98</sup> Government imaginations have been captured by this rush to invest in and integrate AI (including generative AI, but also other AI technologies) into a wide range of government processes and services, in and outside of the military domain. In advance of the first global AI Summit in November 2023, the then prime minister of the UK, Rishi Sunak, announced that the UK would deploy AI tools to clamp down on benefit fraud, in spite of growing evidence of their discriminatory outcomes when used for fraud detection in the UK and elsewhere.<sup>99</sup> In September 2024, the UK’s College of Policing called for AI to be implemented across the police force to automate police paperwork,<sup>100</sup> and the current UK prime minister, Sir Keir Starmer, announced in January 2025 that AI would be “mainlined into the UK’s veins”, with indications that the government would make public data, including anonymised NHS data, available to aid the growth of AI businesses, and use the technology in public service delivery, including in the healthcare system.<sup>101</sup> The UK’s Department of Work and Pensions (DWP) has already announced the rollout of a number of generative AI tools in the sphere of social security. One such tool, known as A-cubed, was fed thousands of guidance documents to subsequently inform work coaches on how to support their claimants into work. Like many of the other generative AI tools announced to be in use by the UK government, A-cubed has since been dropped. The use of these tools to simplify a normally long and involved process risks creating significant errors and introducing automation bias among the DWP workforce.<sup>102</sup>

Similarly, in the USA, then President Joe Biden signed an executive order on *Advancing United States Leadership in Artificial Intelligence Infrastructure* in January 2025, especially in reference to the development and usage of generative AI systems in contexts of “national security, including with respect to logistics, military capabilities, intelligence analysis, and cybersecurity.”<sup>103</sup> This was followed by a broader shift towards deregulation under the current US administration of President Donald Trump, as corporate tech power has garnered unprecedented influence in matters related to governance, and on policies related

<sup>94</sup> Tara Copp, Elizabeth Dwoskin, and Ian Duncan, Anthropic’s AI tool Claude central to U.S. campaign in Iran, amid a bitter feud, 4 March 2026, <https://www.washingtonpost.com/technology/2026/03/04/anthropic-ai-iran-campaign/> (accessed on 14 May 2026).

<sup>95</sup> *Statement from Dario Amodei on our discussions with the Department of War* (no date). Available at: <https://www.anthropic.com/news/statement-department-of-war> (Accessed: 6 March 2026).

<sup>96</sup> *Anthropic officially designated a supply chain risk by Pentagon* (2026). Available at: <https://www.bbc.com/news/articles/cn5g3z3xe65o> (Accessed: 6 March 2026).

<sup>97</sup> *Anthropic officially designated a supply chain risk by Pentagon* (2026). Available at: <https://www.bbc.com/news/articles/cn5g3z3xe65o> (Accessed: 6 March 2026).

<sup>98</sup> John Letzing, “What is ‘sovereign AI’ and why is the concept so appealing (and fraught)?”, 13 November 2024, World Economic Forum, <https://www.weforum.org/stories/2024/11/what-is-sovereign-ai-and-why-is-the-concept-so-appealing-and-fraught/>

<sup>99</sup> Amnesty International, “Global/UK: Artificial Intelligence Summit must not ignore severe rights harms of fraud-detection tech”, 1 November 2023, <https://www.amnesty.org/en/latest/news/2023/11/global-uk-artificial-intelligence-summit-must-not-ignore-severe-rights-harms-of-fraud-detection-tech/>

Amnesty International, *Netherlands: Profiled Without Protection: Students in the Netherlands Hit by Discriminatory Fraud Detection System* (Index: EUR 35/8794/2024), 20 November 2024, <https://www.amnesty.org/en/documents/eur35/8794/2024/en/>

Amnesty International, *Xenophobic Machines: Discrimination Through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal* (Index: EUR 35/4686/2021), 25 October 2021 <https://www.amnesty.org/en/documents/eur35/4686/2021/en/>

Foxglove, “New case: secret algorithm targets disabled people unfairly for benefit probes – cutting off life-saving cash and trapping them in call centre hell”, 1 December 2021, <https://www.foxglove.org.uk/2021/12/01/secret-dwp-algorithm/>

<sup>100</sup> Amy-Clare Martin, “AI needs to be injected into police ‘like heroin into bloodstream’, says top officer”, 10 September 2024, The Independent, <https://www.independent.co.uk/news/uk/crime/artificial-intelligence-police-paperwork-andy-marsh-b2610133.html>

<sup>101</sup> Robert Booth, “‘Mainlined into UK’s veins’: Labour announces huge public rollout of AI”, 12 January 2025, The Guardian, <https://www.theguardian.com/politics/2025/jan/12/mainlined-into-uks-veins-labour-announces-huge-public-rollout-of-ai>

<sup>102</sup> Robert Booth, “AI prototypes for UK welfare system dropped as officials lament ‘false starts’”, 27 January 2025, The Guardian, <https://www.theguardian.com/technology/2025/jan/27/ai-prototypes-uk-welfare-system-dropped>; Amnesty International, “UK: government’s unchecked use of tech and AI systems leading to exclusion of people with disabilities and other marginalized groups”, 10 July 2025, <https://www.amnesty.org/en/latest/news/2025/07/uk-governments-unchecked-use-of-tech-and-ai-systems-leading-to-exclusion-of-people-with-disabilities-and-other-marginalized-groups/>

<sup>103</sup> The White House, *Executive Order on Advancing United States Leadership in Artificial Intelligence Infrastructure*, 14 January 2025, <https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2025/01/14/executive-order-on-advancing-united-states-leadership-in-artificial-intelligence-infrastructure/>



to racial justice, refugees' and migrants' rights,<sup>104</sup> reproductive rights<sup>105</sup> and protest.<sup>106</sup> Through the US Department of Government Efficiency (DOGE), under the initial leadership of tech entrepreneur and billionaire Elon Musk, there have been clear indications of a steady rollback of hard-fought human rights protections,<sup>107</sup> alongside substituting AI in place of genuine oversight and accountability.<sup>108</sup>

In addition, authorities in China are quickly adopting DeepSeek R1 in areas ranging from service delivery to the justice system,<sup>109</sup> despite well-documented concerns about the high risks of discriminatory and biased outcomes of the deployment of such systems. China's New Generation AI Development Plan and the Digital Silk Road initiative are likely to be the primary suppliers of AI models across some Asian and African states in particular.<sup>110</sup> Following the launch of the state-backed DeepSeek R1, the US-based tech company Nvidia, the world's largest developer of graphics processing units (GPUs), lost US\$593 billion in value on the New York Stock Exchange,<sup>111</sup> with Microsoft and Alphabet (Google's parent company) following suit, as investors speculated that US AI market dominance would fail against the backdrop of cheaper and more efficient alternatives.

The renewed "hype" and subsequent adoption of AI, especially generative AI, by governments risks putting systems built on problematic design principles at the heart of various public sector functions, as well as in the military domain.

---

<sup>104</sup> Amnesty International USA, "The digital border: migration, technology and inequality", 21 May 2024, <https://www.amnestyusa.org/reports/the-digital-border-migration-technology-and-inequality/>

Amnesty International, *Primer: Defending the Rights of Refugees and Migrants in the Digital Age* (Index: POL 40/7654/2024), 5 February 2024, <https://www.amnesty.org/en/documents/pol40/7654/2024/en/>

<sup>105</sup> Amnesty International UK, "USA: Social media companies' removal of abortion-related content may hinder access to abortion care", 1 June 2024, <https://www.amnesty.org.uk/press-releases/usa-social-media-companies-removal-abortion-related-content-may-hinder-access>

<sup>106</sup> Amnesty International, 21 August 2025, "USA/Global: Tech made by Palantir and Babel Street pose surveillance threats to pro-Palestine student protestors & migrants", <https://www.amnesty.org/en/latest/news/2025/08/usa-global-tech-made-by-palantir-and-babel-street-pose-surveillance-threats-to-pro-palestine-student-protestors-migrants/>

<sup>107</sup> Olga Akselrod and Cody Venzke, "Trump's efforts to dismantle AI protections, explained", 11 February 2025, ACLU, <https://www.aclu.org/news/privacy-technology/trumps-efforts-to-dismantle-ai-protections-explained>

<sup>108</sup> Hannah Natanson and others, "DOGE builds AI tool to cut 50 percent of federal regulations", 26 July 2025, Washington Post, <https://www.washingtonpost.com/business/2025/07/26/doge-ai-tool-cut-regulations-trump/>

<sup>109</sup> Tobin, M. and Fu, C., "From courtrooms to crisis lines, Chinese officials embrace DeepSeek", 18 March 2025, New York Times, <https://www.nytimes.com/2025/03/18/business/china-government-deepseek.html>

<sup>110</sup> MEMRI TV, "Chinese professor Shen Yi's bold vision and its global implications: DeepSeek, the digital silk road, and China's AI gambit", 21 February 2025, <https://www.memri.org/tv/chinese-professor-shen-yi-deepseek-digital-silk-road>

<sup>111</sup> Sinéad Carew and others, "DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss", 28 January 2025, <https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/>

# 5. HUMAN RIGHTS TENSIONS IN GENERATIVE AI DESIGN

Training many modern generative AI systems requires large amounts of data. The easiest method for developers to obtain this training data is to collect it from the web. These data inputs are usually most readily available in English (or the “resourcedness”), meaning the outputs of generative AI models skew towards English (that is, they are fed largely on English language content). They also tend to be dominated by social, cultural and political norms most visible and available online, typically skewing toward western, anglophone, global minority norms.

In our analysis, we find that both the *scale* of this data gathering for training purposes and the representational issues in the training data span multiple international human rights laws and standards, from privacy violations and non-consensual data collection, to discrimination, harassment, and threats to the rights to freedom of expression and freedom of thought. In this chapter, we detail further this human rights analysis.

## 5.1 UNLAWFUL WEB-SCRAPING AND MASS INVASIONS OF PRIVACY BY DESIGN

While there are generative AI models that do not rely on invasive, mass collection of data – for example, small-language models (SLMs) – many popular and publicly available generative AI tools do, and these are the focus of this briefing’s analysis.

Most commercially available large generative AI systems (such as LLMs and other multi-modal models) cannot function without web scraping for training data, which renders aspects of the design of generative AI systems fundamentally at odds with international human rights law (IHL).

Where companies developing generative AI products have policies to protect privacy, these often do not comply with data protection laws and principles and expressly allow unhindered web-scraping for the purposes of generative AI model development. According to their privacy policies, OpenAI, Google, and Meta employ extensive data collection practices to develop their generative AI products. OpenAI collects user interactions, prompts, and feedback from services like ChatGPT to refine model performance, while also utilising vast amounts of publicly available web content for training data through web crawling.<sup>112</sup> Google’s Gemini collects conversation history, location data, feedback, and usage information to personalise experiences and improve AI capabilities, similarly relying on large-scale web scraping and licensed datasets for model training.<sup>113</sup> Meta gathers user content, interactions, and behavioural data across its platforms

---

<sup>112</sup> *Privacy policy* (no date). Available at: <https://openai.com/en-GB/policies/row-privacy-policy/> (Accessed: 26 February 2026).

<sup>113</sup> *Gemini Apps Privacy Hub - Gemini Apps Help* (no date). Available at: <https://support.google.com/gemini/answer/13594961?hl=en> (Accessed: 26 February 2026).

(Facebook, Instagram, Messenger) to enhance AI features, and constructs training datasets through both user-generated content and publicly accessible web data.<sup>114</sup> All three companies engage in web crawling to build massive training data, with unclear processes around how personal data is excluded from this process and the eventual training data, raising significant questions for consent.

OpenAI's GPT model and other GPT-based products depend on web scraping, a process by which automated software extracts data from websites, resulting in a vast yield of data without which none of their systems can operate. Reports indicate that GPT-3, the model used to power the first public version of ChatGPT, was constructed using the Common Crawl web scraper,<sup>115</sup> which at the time included at least 60 million domains scraped over the course of 12 years.<sup>116</sup> While data on the makeup of GPT-5's training data is currently unavailable, it is highly likely that more recent models build on existing and expanded corpora produced through web crawling.<sup>117</sup>

Reports indicate that up to 60% of the training data used to create systems like ChatGPT came from Common Crawl, a web crawler hosting one of the largest available repositories of web-scraped data, ranging from mainstream news sites to social media platforms, including textual and audio-visual data. As researchers have noted, some web crawlers have been known to scrape even the most intimate data, including medical patients' diagnostic imagery.<sup>118</sup>

Meanwhile, 40% of the training data for GPT was reportedly fed by OpenAI engineers using a broad roster of materials including novels, scripts and artwork.<sup>119</sup> Generative AI developers use these input data sources to generate "synthetic" data; that is, imagery, text, videos and audio which reflects aspects of the input data but reorganized in often different and distinct ways. Nevertheless, these purportedly "novel" outputs take on characteristics that are informed by the non-consensual, mass-collected training data, including people's likeness and patterns of writing, speech and behaviour. AI companies and developers of some of the largest and most lucrative AI products depend on this extractive process of mass and indiscriminate collection of training data for the generation of profit.

Google's Gemini multimodal AI model relies on the existing web crawling infrastructure of Google Search,<sup>120</sup> enabling it to access one of the widest ranging bodies of searchable content available on the web — including text, audio, images, video and computer code — as its corpus of training data.<sup>121</sup> In 2023, Google updated its privacy policy to reflect that it would use such data to train its AI models at large. While little is known about the exact composition and source of Gemini's training data, its LLM predecessor – LaMDA – was pre-trained on 1.56 trillion words of public data and web text.<sup>122</sup>

Meta's LLaMa model relies on an internally developed web crawler named Meta External Agent,<sup>123</sup> which was announced to developers in 2024. Early that year, Meta's CEO Mark Zuckerberg reportedly stated that Meta's competitive edge in generative AI would come from "hundreds of billions of publicly shared images and tens of billions of public videos, which we estimate is greater than the common crawl data set" made available through their platforms, including Facebook and Instagram.<sup>124</sup> Meta's use of training data, in excess of what is already an enormous corpus available via Common Crawl, would indicate that it is training its AI models on data shared on its platforms by users, posing significant risks to their right to privacy.<sup>125</sup>

---

<sup>114</sup> Meta Privacy Policy - How Meta collects and uses user data (no date). Available at: <https://www.facebook.com/privacy/policy/> (Accessed: 26 February 2026).

<sup>115</sup> Common Crawl, "Overview", <https://commoncrawl.org/overview> (accessed on 8 January 2026).

<sup>116</sup> Liz O'Sullivan and John P. Dickerson, "Here are a few ways GPT-3 can go wrong" (previously cited).

<sup>117</sup> Baack, S. (2024) "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl", *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '24), pp. 2199–2208. Available at: <https://doi.org/10.1145/3630106.3659033>.

<sup>118</sup> Lauren Leffer, "Your personal information is probably being used to train generative AI models", 19 October 2023, *Scientific American*, <https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/>

<sup>119</sup> Kalpana Tyagi, "Copyright, text & data mining and the innovation dimension of generative AI", 9 March 2024, *Journal of Intellectual Property Law & Practice*, Volume 19, Issue 7, <https://doi.org/10.1093/jiplp/ijpae028>

<sup>120</sup> Jess Weatherbed, "Google admits it's training AI on scraped web data", 5 July 2023, *The Verge*, <https://www.theverge.com/2023/7/5/23784257/google-ai-bard-privacy-policy-train-web-scraping>

<sup>121</sup> Dan Milmo, "Google says new AI model Gemini outperforms ChatGPT in most tests", 6 December 2023, *The Guardian*, <https://www.theguardian.com/technology/2023/dec/06/google-new-ai-model-gemini-bard-upgrade>

<sup>122</sup> Romal Thoppilan and others, "LaMDA: language models for dialog applications", 10 February 2022, *arXiv*, <https://doi.org/10.48550/ARXIV.2201.08239>

<sup>123</sup> Kali Hays, "A new web crawler launched by Meta last month is quietly scraping the web for AI training data", 20 August 2024, *Fortune*, <https://fortune.com/2024/08/20/meta-external-agent-new-web-crawler-bot-scrape-data-train-ai-models-llama/>

<sup>124</sup> Zeff, M. (2024) 'Zuck Brags About How Much of Your Facebook, Instagram Posts Will Power His AI', *Gizmodo*, 5 February. Available at: <https://gizmodo.com/zuck-brags-how-much-facebook-instagram-posts-power-ai-1851225278> (Accessed: 26 February 2026).

<sup>125</sup> Pascale Davies, "Meta is about to use Europeans' social posts to train its AI. Here's how you can prevent it", 13 May 2025, *EuroNews*, <https://www.euronews.com/next/2025/05/13/meta-is-about-to-use-europeans-social-posts-to-train-its-ai-heres-how-you-can-prevent-it>

Stable Diffusion and Midjourney rely on the LAION-5B image database for model training,<sup>126</sup> which contains captions and links of at least 2.3 billion images, many of which have been derived from Common Crawl using an automated filtration system known as CLIP.<sup>127</sup> The sheer scale of this database, and the methods of bulk and wide data collection and capture, introduces significant issues pertinent to data protection, not least the non-consensual reproduction of personally identifiable information linked to artists.<sup>128</sup> Even more worryingly, as revealed in December 2023, LAION—which Stable Diffusion is reliant on to function—<sup>129</sup> appears to have links to Child Sexual Abuse Material (CSAM).<sup>130</sup> This comes on the back of the designers of CLIP, the filtration system used to sift through Common Crawl data, having already stated in their documentation that they did not intend for the product to be deployed for image searching and/or in an unconstrained environment.<sup>131</sup>

There are immense risks associated with this level of mass data collection. In particular, such systems increase the possibility of personal data leakage. In December 2023, researchers demonstrated a vulnerability in ChatGPT that saw it leak sensitive personal data used as training data in its responses to users, when instructed to repeatedly write the word “poem” in the chat interface.<sup>132</sup> Researchers behind these revelations noted that generative AI chatbots can be prompted to unveil gigabytes of training data from some of the most commonly used generative AI models, without any prior knowledge of the training data.<sup>133</sup>

While some developers of generative AI tools, such as Meta and Google, offer “opt-out” clauses for users who do not want their personal data to be used to train generative AI models, such clauses are neither comprehensive nor effective, as they do not exclude data already scraped from the web using proprietary or third-party data scrapers.<sup>134</sup> Many of the most popular generative AI systems to date have been trained on datasets the provenance of which are being withheld or in other ways obfuscated.

The surveillance implications and potential unlawfulness of these products have also been noted by data protection bodies, globally. In January 2024, Italy’s data protection agency formally accused OpenAI of violating Europe’s General Data Protection Regulation (GDPR) through its scraping. OpenAI has claimed that its actions are justified under a public interest exception to GDPR’s requirement of consent for personal data use.<sup>135</sup> Poland’s Personal Data Protection Office is also investigating the company for GDPR violations.<sup>136</sup> In Germany, the Regional Court of Munich found OpenAI models infringed the copyright of Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte (GEMA),<sup>137</sup> a major collecting society for musical works. It ruled that OpenAI models infringed copyright by memorising and reproducing copyright material in outputs, noting that works could easily be reproduced on the basis of simple prompts.<sup>138</sup> In the Netherlands, the data protection authority published a study indicating that chatbots provide distorted and polarised voting advice. The Data Protection Authority (DPA) found that ChatGPT and other companies and products underrepresented centrist parties and directed users towards far-right or green-labour parties. The DPA noted that these systems may qualify as both GPAI models and high-risk systems under the AI Act when used to provide voting advice (this also has implications for the right to freedom of thought, see

<sup>126</sup> John Naughton, Artists may make AI firms pay a high price for their software’s ‘creativity’, 28 October 2023, <https://www.theguardian.com/commentisfree/2023/oct/28/artists-artificial-intelligences-lawsuits-scraping-midjourney-dataset-nightshade> (accessed on 14 May 2026).

<sup>127</sup> Eryk Salvaggio, LAION-5B, Stable Diffusion 1.5, and the Original Sin of Generative AI, 2 January 2024, <https://techpolicy.press/laion5b-stable-diffusion-and-the-original-sin-of-generative-ai> (accessed on 14 May 2026).

<sup>128</sup> Barry Scannell and Leo Moore, Generative AI Generates Infringement Litigation, 19 January 2023, <https://www.williamfry.com/knowledge/generative-ai-generates-infringement-litigation/> (accessed on 14 May 2026).

<sup>129</sup> Eryk Salvaggio, LAION-5B, Stable Diffusion 1.5, and the Original Sin of Generative AI, 2 January 2024, <https://techpolicy.press/laion5b-stable-diffusion-and-the-original-sin-of-generative-ai> (accessed on 14 May 2026).

<sup>130</sup> Davey Alba and Rachel Metz, Large AI Dataset Has Over 1,000 Child Abuse Images, Researchers Find, 20 December 2023, <https://www.bloomberg.com/news/articles/2023-12-20/large-ai-dataset-has-over-1-000-child-abuse-images-researchers-find> (accessed on 14 May 2026).

<sup>131</sup> Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, Multimodal datasets: misogyny, pornography, and malignant stereotypes, 5 October 2021, <http://arxiv.org/abs/2110.01963> (accessed on 14 May 2026).

<sup>132</sup> Lily Hay Newman and Andy Greenberg, “ChatGPT spit out sensitive data when told to repeat ‘poem’ forever”, 2 December 2023, <https://www.wired.com/story/chatgpt-poem-forever-security-roundup/>

<sup>133</sup> Milad Nasr, “Scalable extraction of training data from (production) language models”, 28 November 2023, [arXiv, https://doi.org/10.48550/arXiv.2311.17035](https://arxiv.org/abs/2311.17035)

<sup>134</sup> Gemini Apps Privacy Hub - Gemini Apps Help (no date). Available at: <https://support.google.com/gemini/answer/13594961?hl=en> (Accessed: 26 February 2026); Meta Privacy Policy - How Meta collects and uses user data (no date). Available at: <https://www.facebook.com/privacy/policy/> (Accessed: 26 February 2026).

<sup>135</sup> Natasha Lomas, ChatGPT is violating Europe’s privacy laws, Italian DPA tells OpenAI, 29 January 2024, <https://techcrunch.com/2024/01/29/chatgpt-italy-gdpr-notification/> (accessed on 21 May 2026)

<sup>136</sup> Poland investigates OpenAI over privacy concerns, 21 September 2023, <https://www.reuters.com/technology/poland-investigates-openai-over-privacy-concerns-2023-09-21/> (accessed on 21 May 2026)

<sup>137</sup> Association of creatives representing 100,000 composers, lyricists and publishers.

<sup>138</sup> 2025 - Bayerisches Staatsministerium der Justiz, <https://www.justiz.bayern.de/gerichte-und-behoerden/landgericht/muenchen-1/presse/2025/11.php> (accessed on 14 May 2026).

Chapter 5.5).<sup>139</sup> This takes place in the context of big tech companies, including OpenAI, reportedly investing massively in lobbying in Europe to skirt new regulations.<sup>140</sup>

The bulk and mass collection of training data for generative AI products, through web scraping, is analogous to the type of mass surveillance at the core of more “narrow” AI systems, such as facial recognition technologies for identification (1:n FRT), for which a reference image database of mass and bulk scraped images is required to function. This image database is usually generated without knowledge and consent. Amnesty International has long held that AI-driven surveillance tools, such as FRT, that scan, capture and often store data on massive databases curated without individuals’ knowledge and consent, are tools of mass surveillance by design. With mass web scraping powering most standalone generative AI models, generative AI companies are also including personal data without knowledge and consent, and as such risk committing mass invasions of privacy. While presenting similar human rights risks, generative AI systems compound many of these as the technology operates at a greater scale, both in its wide range of applications and in its supply chains.

At worst, generative AI systems built on invasive, mass and indiscriminate data collection, risk empowering and enabling more expansive unlawful state surveillance, through the integration of new generative AI functionalities into already existing surveillance tools and practices such as the use of FRT to monitor protestors and public spaces.<sup>141</sup> It is because of the design parameters described in this briefing, that generative AI tools can rapidly expand both offline and online surveillance for law enforcement and policing. The data pipeline that is based on mass scale web scraping indirectly provides access to huge amounts of public domain data and the functionalities of generative AI tools themselves could be used, for instance, to make social media monitoring more expansive, or at least incentivise this through its promised capabilities.<sup>142</sup> When deployed in the context of high-risk environments, such as contexts of conflict and warfare, they may even be used to sanitize international crimes behind a veneer of sophisticated automation. For instance, in March 2025, +972 Magazine, Local Call and The Guardian reported on the development of an LLM trained on Arabic conversations obtained through mass surveillance of Palestinians in the Occupied Palestinian Territory by Israeli security forces.<sup>143</sup> Existing LLMs are only capable of processing standard Arabic, rather than spoken dialects; reportedly, developers of the LLM had to resort to the development of their own language model, trained on mass collected Palestinian conversational data, without the speakers’ knowledge and consent.<sup>144</sup> The system allegedly allows for significant “population control”, including the monitoring and tracking of “persons of interest” to the authorities, including human rights activists, conversations between Palestinians, and how people interact with each other and across space.<sup>145</sup>

## 5.2 DATA CENTRES AND ENVIRONMENTAL IMPACT

Despite the seemingly automated and nebulous nature of generative AI tools, in particular against the backdrop of the increasing migration of data to the “cloud”, an often-overseen material input makes up a critical part of its data pipeline, namely, resource-intensive data centre infrastructures that power the computation needed by these systems.

As demands for generative AI tools have increased, the demands on data centres have followed suit, with larger models now requiring higher amounts of processing power, energy and cooling. Data centres – clusters of concentrated, high-performing computers – provide the processing capabilities of AI systems, carrying out the AI process itself across a network. Their massive energy requirements, in addition to requirements for large quantities of water for cooling, have raised the alarm for environmental experts. Google’s own sustainability report from 2024 noted a staggering 48% increase in the company’s greenhouse

---

<sup>139</sup> AP waarschuwt: chatbots geven vertekend stemadvies | Autoriteit Persoonsgegevens,

<https://www.autoriteitpersoonsgegevens.nl/actueel/ap-waarschuwt-chatbots-geven-vertekend-stemadvies> (accessed on 14 May 2026).

<sup>140</sup> Big Tech lobbying is derailing the AI Act | Corporate Europe Observatory, <https://corporateeurope.org/en/2023/11/big-tech-lobbying-derailing-ai-act> (accessed on 14 May 2026).

<sup>141</sup> Amnesty International, Ban the Scan campaign site, <http://banthescan.amnesty.org/index.html> (accessed on 14 May 2026).

<sup>142</sup> 9211, Advances in AI Increase Risks of Government Social Media Monitoring | Brennan Center for Justice, 26 July 2023, <https://www.brennancenter.org/our-work/analysis-opinion/advances-ai-increase-risks-government-social-media-monitoring> (accessed on 14 May 2026).

<sup>143</sup> Yuval Abraham, “Israel developing ChatGPT-like tool that weaponizes surveillance of Palestinians”, 6 March 2025, +972 Magazine, <https://www.972mag.com/israeli-intelligence-chatgpt-8200-surveillance-ai/>

<sup>144</sup> Yuval Abraham, “Israel developing ChatGPT-like tool that weaponizes surveillance of Palestinians” (previously cited).

<sup>145</sup> Amnesty International has previously concluded that the use of AI systems such as remote biometric recognition and autonomous targeting systems reinforces the existing restrictions on freedom of movement and automates Israel’s system of apartheid in the Occupied Palestinian Territory.<sup>145</sup> In the context of Israel’s ongoing genocide in Gaza,<sup>145</sup> the use of these AI systems risks further impeding or preventing the delivery of humanitarian aid and contributing to the dehumanization, violence, extensive destruction of civilian infrastructure and relentless killings of Palestinians.



gas emissions since 2019, attributable to data centre and supply chain emissions.<sup>146</sup> Similarly, Microsoft's emissions increased by 29% between 2020 and 2024, attributable to data centres carrying out AI-supporting processes.<sup>147</sup> Yet, as recently as September 2024, Microsoft, BlackRock, Global Infrastructure Partners (GIP) and MGX formed a consortium in a bid to raise US\$100 billion to build new data centres.<sup>148</sup> Amazon has made statements claiming that it is becoming "water positive by 2030",<sup>149</sup> and alleging to have met 100% of its sustainability goals in the realm of renewable energy.<sup>150</sup> Meanwhile, the pressure group Amazon Employees for Climate Justice has pushed back, pointing out the company's lack of adequate assessment of the impact of growing demands for generative AI on its data centres and web services, and calling for transparency on the company's use of low-quality renewable energy credits.<sup>151</sup>

Around the world, data centres are already being built at the expense of historically marginalized communities facing the brunt of the catastrophic consequences of the infrastructure. Communities in Chile, for instance, are fighting against the construction of a Google data centre in Cerrillos, Santiago, an industrial and residential area which has faced severe droughts for years.<sup>152</sup> Communities in Querétaro, Mexico, a similarly drought-stricken area, are also resisting the construction of Amazon data centres.<sup>153</sup> In the UK, following the government's call to increase investment in AI systems,<sup>154</sup> the fresh water demands of AI infrastructure has led to growing concern of forthcoming water shortages.<sup>155</sup> A February 2025 report by the UK's Royal Academy of Engineering, for example, calls for an "extension of mandatory reporting on AI's energy and water use, carbon emissions and e-waste recycling of data centres [and minimizing] risks to the environment, people, and the economy by managing the resource demands of AI systems".<sup>156</sup> The US state of Arizona reportedly approved an 8% increase in electricity production to support the rapidly growing data centre industry, while rejecting plans to bring power to parts of Navajo Nation land, where Indigenous communities have no access to mains electricity.<sup>157</sup> Such cases highlight how the accelerating demands and hype around generative AI risk exacerbating electricity shortages and contributing to unequal distribution of services, displacing appropriate investment and care for historically marginalized, and in particular Indigenous, communities.

In addition to the high environmental risks posed by the operational aspects of data centres, the graphical processing units (GPUs) powering the computers within data centres each come with compounding supply chain issues. For instance, the world's largest supplier of GPUs, NVIDIA, relies on several other companies to supply and partly produce components of their semi-conductor chips based on elements including silicon, copper, gold, tantalum, palladium, cobalt and boron.<sup>158</sup> Many of these elements require engagement in and expansion of mining practices that Amnesty International has already investigated for severe human rights abuses and environmental degradation.<sup>159</sup>

<sup>146</sup> Google, *Environmental Report 2024*, 2024, <https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf>

<sup>147</sup> Microsoft, *2024 Environmental Sustainability Report*, 2024, <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Microsoft-2024-Environmental-Sustainability-Report.pdf>; Dara Kerr, "Google and Microsoft report growing emissions as they double-down on AI", 12 July 2024, NPR, <https://www.npr.org/2024/07/12/g-s1-9545/ai-brings-soaring-emissions-for-google-and-microsoft-a-major-contributor-to-climate-change>

<sup>148</sup> Emil Sayegh, "The billion-dollar AIAI gamble: data centers as the new high-stakes game", 30 September 2024, Forbes, <https://www.forbes.com/sites/emilsayegh/2024/09/30/the-billion-dollar-ai-gamble-data-centers-as-the-new-high-stakes-game/>

<sup>149</sup> Thomas Graham, "Mexico's datacentre industry is booming – but are more drought and blackouts the price communities must pay?", 25 September 2024, The Guardian, <https://www.theguardian.com/global-development/2024/sep/25/mexico-datacentre-amazon-google-queretaro-water-electricity>

<sup>150</sup> Pablo Jiménez Arandía, "Deciphering AI water consumption: how Amazon hides how much water its cloud consumes in Spain", 7 March 2025, El País/Pulitzer Center, [https://pulitzercenter.org/stories/deciphering-ai-water-consumption-how-amazon-hides-how-much-water-its-cloud-consumes-spain?utm\\_source=twitter&utm\\_medium=social&utm\\_campaign=AI-Accountability](https://pulitzercenter.org/stories/deciphering-ai-water-consumption-how-amazon-hides-how-much-water-its-cloud-consumes-spain?utm_source=twitter&utm_medium=social&utm_campaign=AI-Accountability)

<sup>151</sup> Amazon Employees for Climate Justice, *Burns Trust: The Amazon Unsustainability Report*, published June 2024, updated October 2024, <https://static1.squarespace.com/static/65681f099d7c3d48feb86a5f/t/6721c4047213ea343e50536f/1730266118471/unsustainability-report-2.pdf>

<sup>152</sup> Sebastián Lehuédé, "Big Tech's new headache: data centre activism flourishes across the world", 2 November 2022, Media@, <https://blogs.lse.ac.uk/mediasec/2022/11/02/big-techs-new-headache-data-centre-activism-flourishes-across-the-world/>

<sup>153</sup> <https://www.theguardian.com/global-development/2024/sep/25/mexico-datacentre-amazon-google-queretaro-water-electricity>

<sup>154</sup> UK Government, "Prime Minister sets out blueprint to turbocharge AI", 13 January 2025, <https://www.gov.uk/government/news/prime-minister-sets-out-blueprint-to-turbocharge-ai>

<sup>155</sup> Zoe Kleinman and Brian Wheeler, "Concern the UK's AI ambitions could lead to water shortages", 7 February 2025, BBC News, <https://www.bbc.com/news/articles/ce85wx9ijndo>

<sup>156</sup> Royal Academy of Engineering, "UK Government urged to promote, prioritise and invest in sustainable AI to become global leader in AI frugality and efficiency", 7 February 2025, <https://raeng.org.uk/news/uk-government-urged-to-promote-prioritise-and-invest-in-sustainable-ai-to-become-global-leader-in-ai-frugality-and-efficiency>

<sup>157</sup> The Washington Post, "Amid Arizona's data center boom, many Native Americans live without power" 23 December 2024, <https://www.washingtonpost.com/technology/2024/12/23/arizona-data-centers-navajo-power-aps-srp/>

<sup>158</sup> Ana Valdivia, "The supply chain capitalism of AI" (previously cited).

<sup>159</sup> Amnesty International, "Forced evictions at industrial cobalt and copper mines in the DRC", 12 September 2023, <https://www.amnesty.org/en/latest/news/2023/09/drc-cobalt-and-copper-mining-for-batteries-leading-to-human-rights-abuses/>

While generative AI products lend themselves to fantasies of frictionless futures void of material concerns, a deeper interrogation of their supply chains unveils the technology's deep ties with forms of environmental exploitation that are endemic to the AI industry. As demands for larger generative AI models increase, so too do the requirements for processing power, which adds further strain to the supply of essential minerals for semi-conductor chips, energy, cooling and land for data centres.

## 5.3 PERPETUATING STEREOTYPES, BIASES AND DISCRIMINATION

Because generative AI products are created by distilling an “average” view from colossal tranches of data, absorbing whatever biases, stereotypes and gaps in understanding they present, such systems can have the effect of glossing over minority views due to their under-representation in training data and uncritically parroting dominant narratives. In other words, in societies where systemic racism prevails, the “average” view distilled by generative AI will reflect racist assumptions.

This is most striking when generative AI tools infer and repeat discriminatory views through stereotypes and prejudices, such as making an Asian person look white when asked to turn their selfie into a “professional headshot,”<sup>160</sup> or instructing a user, in response to a question determining whether someone should be tortured, to check their country of origin.<sup>161</sup>

As the demand for scaling multi-modal AI models increases, so too do the associated problems of bias and discrimination. A 2023 study led by Abeba Birhane demonstrates that, as datasets powering AI models scale up, the presence of hateful and discriminatory content also increases, along with negative stereotypes and prejudices, including for example the association of Black people's faces with “criminal” attributes.<sup>162</sup> The study furthermore concluded that there was “evidence of hateful, aggressive, and targeted content in the alt-text [textual descriptions of images included in training datasets] audit and evidence of racist stereotyping and dehumanizing classification in the models, particularly towards Black men, all of which exacerbates with dataset size”.<sup>163</sup> In other words, racial bias permeates training data, the AI model itself, and its outputs, with the model and outputs becoming more hateful and racist as the size of the model (and training dataset) increases.

In one Australian study, researchers prompted OpenAI's generative art tool, DALL-E, to generate individual and group depictions of Australian pharmacists.<sup>164</sup> The study notes that 64% of pharmacists in Australia are women, yet in the DALL-E outputs, “69.7% of pharmacists were depicted as men, 29.7% as women, 93.5% as a light skin tone, 6.5% as mid skin tone, and 0% as dark skin tone”.<sup>165</sup> Similarly, researchers at Stanford University found that images generated through Stable Diffusion (an image generation tool developed by Stability AI) and OpenAI's DALL-E returned racist and sexist renderings, including through associating prompts like the word “Africa” with poverty, or “poor” with darker complexion.<sup>166</sup> An MIT Technology Review investigation from October 2025 furthermore revealed that OpenAI's GPT-5 is rife with caste bias, and that this remains unaddressed.<sup>167</sup>

A 2024 UNESCO study on OpenAI's GPT and Meta's LLaMa highlighted significant gender bias, homophobia and racial stereotyping. Notably, women “were described as working in domestic roles far more often than men – four times as often by one model – and were frequently associated with words like ‘home’, ‘family’ and ‘children’, while male names were linked to ‘business’, ‘executive’, ‘salary’, and ‘career’”.<sup>168</sup>

---

<sup>160</sup> Sawdah Bhaimiya, “An Asian MIT student asked AI to turn an image of her into a professional headshot: It made her white, with lighter skin and blue eyes”, 1 August 2023, Business Insider, <https://www.businessinsider.com/student-uses-playground-ai-for-professional-headshot-turned-white-2023-8?r=US&IR=T>

<sup>161</sup> @spiantado, X post, 4 December 2022, <https://x.com/spiantado/status/1599462405225881600?s=20>

Piantadosi prompted ChatGPT to write a series of Python programmes for making decisions.

<sup>162</sup> Abeba Birhane and others, “On hate scaling laws for data-swamps”, 28 June 2023, arXiv, <https://doi.org/10.48550/ARXIV.2306.13141>

<sup>163</sup> Abeba Birhane and others, “On hate scaling laws for data-swamps” (previously cited).

<sup>164</sup> Geoffrey Currie and others, “Gender and ethnicity bias in generative artificial intelligence text-to-image depiction of pharmacists” (previously cited).

<sup>165</sup> Geoffrey Currie and others, “Gender and ethnicity bias in generative artificial intelligence text-to-image depiction of pharmacists” (previously cited).

<sup>166</sup> Ananya, “AI image generators often give racist and sexist results: can they be fixed?” (previously cited).

<sup>167</sup> Nilesh Christopher, OpenAI is huge in India. Its models are steeped in caste bias., 1 October 2025, <https://www.technologyreview.com/2025/10/01/1124621/openai-india-caste-bias/> (accessed on 14 May 2026)

<sup>168</sup> UNESCO, “Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes”, 7 March 2024, <https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes>

Bias can also be more subtle. Dominant media narratives and amplified online perspectives are baked into the training data used for generative AI systems, and can perpetuate unjust, inaccurate and harmful perspectives that risk contributing to the erasure of Indigenous Peoples and other marginalized or oppressed groups' knowledge, heritage and representation. For example, researcher and artist Ameera Kawash used the generative art tool Midjourney to generate hyper-realistic photos of cats wearing Keffiyehs (a symbol of Palestinian identity) in different areas of occupied East Jerusalem. The tool was unable to place cats in particular places, including the Dome of the Rock – the site of Al-Aqsa Mosque – where the tool would only generate images of cats wearing skullcaps (a symbol of Jewish identity).<sup>169</sup> The generative AI system was, in other words, unable to transcend the biased training data that contained particular preconceptions about cultural heritage and belonging in Jerusalem, which excluded Palestinians.

In September 2024, two tech companies, Intel and VMware, presented a project leveraging generative AI systems for the enhancement of “surveillance, retail management, and public safety”,<sup>170</sup> promising to improve scene-by-scene analysis, anomaly detection and other video analysis with “minimal human intervention”, using video-based LLMs.<sup>171</sup> Increasingly used together with other remote biometric surveillance systems, such as facial recognition, companies claim that generative AI models such as GANs can generate synthetic facial data to aid the facial recognition algorithms where there are partial or incomplete matches based on low-resolution input imagery.<sup>172</sup>

This “artificial” enhancement of the input imagery relies on approximating missing data using synthetic data generated from training data, further fuelling bias and discrimination in biometric surveillance systems.<sup>173</sup> For instance, it can lead to false identification in biometric recognition systems, the perpetuation of stereotypes in image generation tools, and the wrongful tagging of content for suppression and removal in automated content-moderation systems. Synthetic data substitutes for incomplete datasets have long been criticized as an AI practice especially prone to biased outputs, under the veneer of accuracy and convenience, as such data already reflects inherent biases. In other words, using synthetic data involves training biased AI models with biased data from other biased AI models, multiplying the problem.<sup>174</sup> Furthermore, new research has unveiled how AI models trained recursively on growing bodies of synthetic/AI generated data collapse over time,<sup>175</sup> leading to outputs that are increasingly distorted and nonsensical.<sup>176</sup> Amnesty International has previously documented the impacts on communities subjected to mass and discriminatory forms of biometric surveillance.<sup>177</sup>

These reproductions of bias and discrimination are also entwined with consent issues. A prominent feature of generative AI systems is their use to represent, mimic or recreate persons and who are often unaware of these outputs and outcomes.<sup>178</sup> In 2023 a 14-year-old in the USA made the news when she wrote to lawmakers about her classmates circulating fake, AI-generated sexually explicit images of her.<sup>179</sup> Similarly, non-consensual sexual imagery ostensibly showing more than 30 local women and girls, created with an “undressing app”, was distributed in a Spanish town.<sup>180</sup> AI-generated and AI-manipulated child sexual abuse material amounts to a small percentage of all reported CSAM. However, these figures seem to accelerate with the increased user numbers of GenAI products. The National Center for Missing & Exploited Children (NCMEC) reported a 1,325% increase in CyberTipline reports involving AI CSAM in 2024.<sup>181</sup> The violence is distinctly gendered and discriminatory in nature as AI-generated CSAM has thus far been found to overwhelmingly portray girls.<sup>182</sup> Targets of AI-driven CSAM, like those of all types

<sup>169</sup> Ameera Kawash, “What a cat in a keffiyeh reveals about AI’s anti-Palestinian bias” (previously cited).

<sup>170</sup> Intel and VMware, *Leveraging Generative AI to Enhance Surveillance, Retail Management, and Public Safety*, 2024, <https://builders.intel.com/solutionslibrary/leveraging-generative-ai-to-enhance-surveillance-retail-management-and-public-safety>

<sup>171</sup> Intel and VMware, *Leveraging Generative AI to Enhance Surveillance, Retail Management, and Public Safety* (previously cited).

<sup>172</sup> M. B. Shahbakhsh and H. Hassanpour, “Empowering face recognition methods using a GAN-based single image super-resolution network”, October 2022, *International Journal of Engineering*, Volume 35, Issue 10, <https://doi.org/10.5829/IJE.2022.35.10A.05>

<sup>173</sup> Alan Blackwell and others, “Computer says ‘don’t know’ - interacting visually with incomplete AI models”. In Sandra Fan and others (eds), *Proceedings of the Workshop on Designing Technologies to Support Human Problem Solving*, University of Washington, <https://www.cl.cam.ac.uk/~afb21/publications/DTSHPS18-Blackwell.pdf>, pp. 5-14.

<sup>174</sup> Danielle Shanley and others, “Getting real about synthetic data ethics: Are AI ethics principles a good starting point for synthetic data ethics?”, 22 February 2024, *EMBO Reports*, Volume 25, Issue 5, <https://doi.org/10.1038/s44319-024-00101-0>

<sup>175</sup> Ilia Shumailov, “AI models collapse when trained on recursively generated data”, 24 July 2024, *Nature*, Volume 631, Issue 8022, <https://doi.org/10.1038/s41586-024-07566-y>

<sup>176</sup> Bernard Marr, “Why AI models are collapsing and what it means for the future of technology”, 19 August 2024, *Forbes*, <https://www.forbes.com/sites/bernardmarr/2024/08/19/why-ai-models-are-collapsing-and-what-it-means-for-the-future-of-technology/>

<sup>177</sup> See: <https://banthescan.amnesty.org/>

<sup>178</sup> Katharine Miller, “Privacy in an AI era: how do we protect our personal information?”, 18 March 2024, <https://hai.stanford.edu/news/privacy-ai-era-how-do-we-protect-our-personal-information>

<sup>179</sup> Tate Ryan-Mosley, “A high school’s deepfake porn scandal is pushing US lawmakers into action”, 1 December 2023, *MIT Technology Review*, <https://www.technologyreview.com/2023/12/01/1084164/deepfake-porn-scandal-pushing-us-lawmakers/>

<sup>180</sup> Emmanuelle Saliba and others, “Mobile apps fueling AI-generated nudes of young girls: Spanish police”, 2 October 2023, *ABC News*, <https://abcnews.go.com/US/mobile-apps-fueling-ai-generated-nudes-young-girls/story?id=103563734>

<sup>181</sup> NCMEC, *NCMEC releases new data: 2024 in numbers*, 8 May 2025.

<sup>182</sup> IWF, *What has changed in the AI CSAM landscape?*, 2024, p. 23.

of CSAM, may experience feelings of shame, humiliation, violation, and a loss of control over their identity. Children depicted in such material might withdraw socially, face bullying and extortion, and in some cases develop anxiety, depression, or self-harm ideation.<sup>183</sup> Platforms have taken little action to prevent the generation and dissemination of such content. For instance, on 25 March 2026, a jury in the Los Angeles superior court ruled that Meta had “violated state consumer protections and misleading parents about the safety of its apps,” and was subsequently ordered to pay \$375 million in civil damages.<sup>184</sup> Witnesses at the trial reportedly testified that a large number of junk reports, generated by an over-reliance on AI moderation, made it impossible to investigate crimes on the platform, including the distribution of CSAM.<sup>185</sup>

In January 2024, singer-songwriter Taylor Swift became one of the most high-profile public figures known to have been subjected to non-consensual explicit image generation using AI, gathering 47 million views before the image posted to X was taken down, or the user was blocked,<sup>186</sup> while cases involving non-consensual explicit image generation of J-pop and K-pop celebrities have also risen dramatically.<sup>187</sup> This application of generative AI tools can thus amount to technology-facilitated gender-based violence.<sup>188</sup>

## 5.4 SHRINKING CIVIC SPACE

Generative AI tools risk censoring protected expression, which can create particularly chilling effects both online and offline in a number of ways, especially due to the limitations of their language resourcedness, leading to discriminatory outcomes. Most available LLMs are “resourced” predominantly in English language content and western and/or Anglocentric culture because they rely on mass-scale web scraping, and do not function well outside of these linguistic contexts, meaning they risk inaccurately or falsely parsing non-western, non-English content and expression as offensive, violent or otherwise problematic. As a result of generative AI companies’ insistence on scale, the fundamentally expansive scope of data capture in their model lacks the function- and -language-specificity needed to produce outputs that can operate in linguistically and culturally diverse contexts. As such, many generative AI systems are highly unpredictable, inaccurate, invasive, biased and resource intensive.

Recently, several investigations have demonstrated that LLMs are susceptible to “censorship tuning” – alignment of the LLM with certain political positions. The Chinese AI model DeepSeek R1 reportedly refuses to return results when prompted with questions regarding the 1989 Tiananmen Square massacre in Beijing, simply stating “sorry, that’s beyond my current scope. Let’s talk about something else”.<sup>189</sup> Scholars have developed approaches, including “refusal discovery”, to query LLMs on topics where the LLM is likely to refuse answering prompts directly. One study found that, in the case of DeepSeek R1, such topics included among others:

**“1. Any content that the Chinese Government may regard as sensitive or restrictive. 2. Any criticism of the Chinese political system. 3. Any speech that may be seen as challenging the Chinese Communist Party. 4. Any negative criticism that might involve Chinese leaders. 5. Any reference to sensitive historical events such as**

<sup>183</sup> For a summary of the impact of real CSAM on children, see Parti/Szabo, *The Legal Challenges of Realistic and AI-Driven Child Sexual Abuse Material: Regulatory and Enforcement Perspectives in Europe*, Laws 2024 (Vol.13), pp. 3-4.

<sup>184</sup> Belanger, A. (2026) *Meta loses trial after arguing child exploitation was “inevitable” on its apps*, *Ars Technica*. Available at: <https://arstechnica.com/tech-policy/2026/03/meta-loses-trial-after-arguing-child-exploitation-was-inevitable-on-its-apps/> (Accessed: 25 March 2026).

<sup>185</sup> McQue, K. (2026) ‘Meta ordered to pay \$375m after being found liable in child exploitation case’, *The Guardian*, 24 March. Available at: <https://www.theguardian.com/technology/2026/mar/24/meta-new-mexico-jury> (Accessed: 25 March 2026).

<sup>186</sup> Kate Conger and John Yoon, “Explicit deepfake images of Taylor Swift elude safeguards and swamp social media” (previously cited).

<sup>187</sup> SCMP, “Japan arrests man in first AI porn case involving deepfakes of J-pop idols and actresses”, 17 October 2025, <https://www.scmp.com/week-asia/people/article/3329378/japan-arrests-man-first-ai-porn-case-involving-deepfakes-j-pop-idols-and-actresses>

E.J. Dickson, “Deepfake porn is still a threat, particularly for K-pop stars”, 7 October 2019, *Rolling Stone*, <https://www.rollingstone.com/culture/culture-news/deepfakes-nonconsensual-porn-study-kpop-895605/>

<sup>188</sup> Amnesty International, *Human Rights Implications of Technology-Facilitated Gender-Based Violence: Submission to the Human Rights Council Advisory Committee* (Index: IOR 40/9284/2025), 24 April 2025, <https://www.amnesty.org/en/documents/ior40/9284/2025/en/>

<sup>189</sup> Donna Lu, “We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan”, 28 January 2025, *The Guardian*, <https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan>

**Tiananmen events,<sup>190</sup> 64 events,<sup>191</sup> Xinjiang Re-education Camp.<sup>192</sup> 6. Any that may involve Taiwan, Tibet, Hong Kong, New Zealand. 8. Conspiracy theories".<sup>193</sup>**

As the use of AI models expands for some of the most basic digital tasks, especially where the scale of its training data lends it a legitimacy for general information sourcing, for instance through integration into search engines and for everyday queries – the censorship of politicized content that skews the historical record sets a dangerous precedent for normalizing censorship and threatens the right to seek information.

Another area in which generative AI systems risk censoring free expression is in their application to content moderation practices by social media companies. LLMs have been adopted rapidly to this end, as a purportedly low-cost means of what has conventionally been a heavily human-driven, yet machine-aided, process.<sup>194</sup> Traditionally, automated content moderation has relied on older basic language models developed in the broader field of natural language processing (NLP),<sup>195</sup> which relied on rules-based systems to filter and detect problematic keywords, on which a model would have been trained. With the new generation of LLMs promising the ability to parse context and detect violating speech beyond keywords, they have rapidly been adopted for automated content moderation on social media platforms.<sup>196</sup> The over-reliance on automated systems that promise faster and more accurate natural language processing through LLMs for content moderation purposes can lead to over-broad censorship of content by historically marginalized communities in particular. Already in 2020, Meta, for example, announced that it was using “super-efficient AI models to detect hate speech”.<sup>197</sup> In February 2024 the company announced that it had started training LLMs on Meta’s community standards,<sup>198</sup> in a bid to automate multi-modal content flagging,<sup>199</sup> and to fast-track content review queues where the model is confident that the content in question does not, in fact, violate a company policy. In the same year, X also announced an increasing shift to machine learning tools for content moderation, in addition to human reviewers.<sup>200</sup> In the lead up to these announcements, the two major social media companies each laid off thousands of content moderation staff.<sup>201</sup> Similarly, in a letter dated 7 November 2025 in response to the UK House of Commons’ Innovation and Technology Committee, TikTok reported that increased sophistication in AI model training meant a reduction in Trust & Safety staff, including staff whose roles involved training content moderators on the TikTok community guidelines.<sup>202</sup>

When used to support content moderation, LLM-based generative AI tools risk flagging, removing, suppressing or otherwise censoring content in Global Majority languages and differing cultural contexts without justification.<sup>203</sup> As tools that are highly limited in their ability to parse context, platforms relying on automated content moderation have already been scrutinized for the repressive consequences of such unequal outcomes. In October 2023, Meta apologized for inserting the word “terrorist” into Instagram profile

<sup>190</sup> Amnesty International, 30 May 2025, “What is the Tiananmen crackdown?”,

<https://www.amnesty.org/en/latest/campaigns/2025/05/what-is-the-tiananmen-crackdown/>

<sup>191</sup> Common reference to bypass censorship on Tiananmen.

<sup>192</sup> Amnesty International, 24 September 2018, “Up to one million detained in China’s mass ‘re-education’ drive”,

<https://www.amnesty.org/en/latest/news/2018/09/china-up-to-one-million-detained/>

<sup>193</sup> Can Rager and others, “Discovering forbidden topics in language models”, 23 May 2025, <https://arxiv.org/html/2505.17441v1>

<sup>194</sup> Oversight Board, “Content moderation in a new era for AI and automation”, <https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/> (accessed on 5 March 2025); Anisha Sircar, “X’s latest content findings reveal troubling trends in AI moderation”, 18 October 2024, Forbes, <https://www.forbes.com/sites/anishasircar/2024/10/18/xs-latest-content-findings-reveal-troubling-trends-in-ai-moderation/>

Tao Huang, “Content moderation by LLM: from accuracy to legitimacy”, <https://arxiv.org/html/2409.03219v2> (accessed 8 January 2026);

Justin Hendrix, “Considering the human rights impacts of LLM content moderation”, 6 July 2025, <https://www.techpolicy.press/considering-the-human-rights-impacts-of-llm-content-moderation/>

European Center for Not-for-Profit Law, *Algorithmic Gatekeepers: The Human Rights Impacts of LLM Content Moderation – Executive Summary*, April 2025, [https://ecnl.org/sites/default/files/2025-04/ECNL\\_LLM\\_CM\\_Executive%20Summary\\_2025.pdf](https://ecnl.org/sites/default/files/2025-04/ECNL_LLM_CM_Executive%20Summary_2025.pdf)

<sup>195</sup> Ghaseminejad Raeini, M. (2025) ‘The evolution of language models: From N-Grams to LLMs, and beyond’, *Natural Language Processing Journal*, 12, p. 100168. Available at: <https://doi.org/10.1016/j.nlp.2025.100168>.

<sup>196</sup> Ghaseminejad Raeini, M. (2025) ‘The evolution of language models: From N-Grams to LLMs, and beyond’, *Natural Language Processing Journal*, 12, p. 100168. Available at: <https://doi.org/10.1016/j.nlp.2025.100168>.

<sup>197</sup> Meta, “How Facebook uses super-efficient AI models to detect hate speech”, 19 November 2020, <https://ai.meta.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/>

<sup>198</sup> Meta, “How Facebook uses super-efficient AI models to detect hate speech” (previously cited).

<sup>199</sup> Tomas Apodaca and Natasha Uzcátegui-Liggett, “How automated content moderation works (even when it doesn’t)”, 1 March 2024,

<https://themarkup.org/automated-censorship/2024/03/01/how-automated-content-moderation-works-even-when-it-doesnt-work>

<https://themarkup.org/automated-censorship/2024/03/01/how-automated-content-moderation-works-even-when-it-doesnt-work>

<sup>200</sup> X, *Global Transparency Report*, 2024, <https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf>

<sup>201</sup> Kari Paul, “Reversal of content policies at Alphabet, Meta and X threaten democracy, warn experts”, 7 December 2023, The Guardian,

<https://www.theguardian.com/media/2023/dec/07/2024-elections-social-media-content-safety-policies-moderation>

<sup>202</sup> Letter from Ali Law, Director of Public Policy and Government Affairs, Northern Europe, TikTok, to Dame Chi Onwurah MP, Chair of Science, Innovation and Technology Committee, dated 7 November 2025,

<https://committees.parliament.uk/publications/50179/documents/270768/default/>

<sup>203</sup> Spandana Singh, “Everything in moderation”, 22 July 2019, New America, <http://newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/>

translations containing the words “Palestinian” and “Alhamdulillah” (which means Praise be to God),<sup>204</sup> and the Palestinian flag emoji.<sup>205</sup>

Social media companies have come to rely heavily on AI systems for content moderation. When paired with a reduction in human capacity to perform content moderation, and the issues in the training data detailed in this briefing, this poses significant risks to historically marginalized and racialized communities. As discussed above, AI systems often reproduce pre-existing societal biases. In the case of content moderation AI,<sup>206</sup> this means that the communities most affected include those directly threatened by ongoing conflict, including those facing genocide, as is the case with Palestinians in Gaza.<sup>207</sup>

Similarly, where generative AI-based content-moderation efforts are used to automatically identify and remove content, or to flag, suppress or “shadow-ban” (partial blocking and removal) particular users, there is a heightened risk to protected speech and expression online. Such instances have been on the rise in recent years,<sup>208</sup> leading, in December 2024, to the Council of Europe calling for human content moderators to remain a cornerstone of content moderation online.<sup>209</sup>

## 5.5 AUTOMATION BIAS AND MANIPULATION

As a technology that mimics human capabilities, human-computer interaction with generative AI models is likely to involve system deference, through the phenomenon that Meredith Broussard has referred to as “techno-chauvinism”; the idea that technology can solve complex social, political and economic issues *better* than human beings.<sup>210</sup> Even prior to the popular availability of generative AI tools, Virginia Eubanks demonstrated in 2018 how case workers who were prompted to use much “narrower” algorithmic tools to assess risks to children known to child protection services in the USA were finding themselves compelled to rely on the visual risk indicators provided to them by the system, rather than their own expertise and judgement.<sup>211</sup> Virginia Eubanks demonstrated how this form of system deference, also referred to as automation bias, led to devastating outcomes for families facing unfair and unjust disciplinary consequences from authorities.

In the context of generative AI systems, this phenomenon risks being reinforced by the false perception that the larger model and training data is equivalent to greater accuracy. Rather, users of more complex and more widely available AI systems, such as chat-based generative AI tools, are at particular risk of accepting generated outputs as reliable, given the increasing scale and centrality of generative AI products to digital querying, paving the way for an inhibition of critical faculties, or at worst, deliberate manipulation.

Research in cognitive science, most recently by academics Celeste Kidd and Abeba Birhane, found that frequency of exposure to fabricated information “predicts how deeply ingrained the belief in that information becomes”.<sup>212</sup> Similarly, studies find that repeated exposure to algorithmic biases has a similar effect; in other words, the millions of generative AI outputs making their way across our screens subtly shape our attitudes to reflect the gaps and biases coded into the system through partial and otherwise manipulated training data.

In late 2024, research by academics Yaqub Chaudhary and Jonnie Penn found that generative AI tools were already being used “to elicit, infer, collect, record, understand, forecast, and ultimately manipulate, modulate, and commodify human plans and purposes, both mundane (e.g., selecting a hotel) and profound (e.g., selecting a political candidate)”.<sup>213</sup> In addition to exposure to fabricated information or biased outputs, the study suggested that even more quotidian aspects of LLM-based technology, such as predictive

---

<sup>204</sup> Samantha Cole, Instagram ‘Sincerely Apologizes’ For Inserting ‘Terrorist’ Into Palestinian Bio Translations, 19 October 2023, <https://www.404media.co/instagram-palestinian-arabic-bio-translation/> (accessed on 21 May 2026)

<sup>205</sup> Amnesty International, “Global: Social media companies must step up crisis response on Israel-Palestine as online hate and censorship proliferate”, 27 October 2023, <https://www.amnesty.org/en/latest/news/2023/10/global-social-media-companies-must-step-up-crisis-response-on-israel-palestine-as-online-hate-and-censorship-proliferate/>

<sup>206</sup> New America, “The limitations of automated tools in content moderation”, <https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation/> (accessed 9 January 2026).

<sup>207</sup> New America, “The limitations of automated tools in content moderation” (previously cited).

<sup>208</sup> Human Rights Watch, *Meta’s Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook*, 21 December 2023, <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and>

<sup>209</sup> Council of Europe, “Regulating content moderation on social media to safeguard freedom of expression”, 4 December 2024, <https://rm.coe.int/as-cult-regulating-content-moderation-on-social-media-to-safeguard-fre/1680b2b162>

<sup>210</sup> Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, 2018.

<sup>211</sup> Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, 2018.

<sup>212</sup> Celeste Kidd and Abeba Birhane, “How AI can distort human beliefs”, 22 June 2023, *Science*, Volume 380, Issue 6651, <https://doi.org/10.1126/science.adi0248>

<sup>213</sup> Chaudhary, Y. and Penn, J. (2024) ‘Beware the Intention Economy: Collection and Commodification of Intent via Large Language Models’, *Harvard Data Science Review* [Preprint], (Special Issue 5). Available at: <https://doi.org/10.1162/99608f92.21e6bbaa>

suggestion, “may interrupt individual thought processes of users, who may subsequently change their views during text composition”.<sup>214</sup> While this may not be malicious, it still puts the power to alter human intention in the hands of developers of LLMs, and ultimately presents a critical risk to freedom of thought. Scholars have long warned that a reliance on such technology for decision-making could atrophy capabilities for moral reasoning.<sup>215</sup>

In other words, as the rapid evolution and dissemination of synthetic content far outpace our ability to critically discern it, there is a heightened risk of susceptibility to disinformation or misinformation.<sup>216</sup> This not only risks reinforcing dominant narratives and tropes as outlined in Chapter 5.3, but even further limits the “range” of expression available to some users, as particular racial, ethnic and cultural realities are denied in favour of dominant narratives, significantly hampering their right to freedom of expression (see Chapter 5.4).

## 5.6 CONCLUSION

Many of the fundamental design features of the generative AI products discussed within this briefing, in particular those pertaining to OpenAI, Google, DeepSeek, Meta, Midjourney, and Stable Diffusion appear incompatible with aspects of international human rights law and standards. The design principles of these products involve massive data collection without consent, biased training processes and the potential for manipulative outputs. As a result, generative AI tools risk mirroring and amplifying many of the problems seen in narrow AI systems with unavoidable implications for the rights to privacy, equality and non-discrimination, and freedom of expression and thought.

This is only further animated by the increasing human rights risks associated with how these systems are deployed and used. Without significant changes to how these systems are developed and deployed, they will continue to undermine international human rights laws and standards. Whether proprietary or open source, these systems require urgent attention to establish clear frameworks for accountability and human rights protection before their impacts become more deeply entrenched in society. Standalone generative AI systems identified to have been built on unlawful web-scraping, or which have demonstrated patterns of discrimination, or both, present significant human rights risks, and in many cases, abuses. Amnesty International contacted DeepSeek, Google, OpenAI, Meta, Stability AI, Intel, VMware, Midjourney, Microsoft, and Amazon, for their responses to findings and concerns regarding standalone generative AI systems and their human rights risks. At the time of publication, OpenAI, Meta, Microsoft, Intel, and Amazon have responded.

Amnesty International sent a letter to OpenAI on 1 April 2026, which received a response on 15 April 2026. In their response, OpenAI mention that their ‘Supplier Code of Conduct’, ‘Model Spec’, and ‘Usage Policies’, provide human rights standards, privacy protection, and prohibits the use of their products in ways which interfere with human rights, and mentions that they have taken steps to “limit the processing of personal information during training, such as by excluding sources that aggregate large amounts of personal data and training our models to avoid responding to requests for private or sensitive information about individuals.” OpenAI in their response also cited their efforts in relation to preserving and protecting rights holders such as vulnerable groups in the context of covert influence activity by state authorities. Amnesty International has not verified such claims nor do they directly impact the findings set out in this publication.

Amnesty International also sent a letter to Meta on 1 April 2026, receiving a response on 21 April 2026. In their response, Meta note their commitment to the UN Guiding Principles, and the OECD Principles on Artificial Intelligence. Meta noted in their response that, given time constraints, they were “unable to provide [...] detailed responses to each portion of your analysis,” followed by a list of links to annual human rights reports (2023-2025), responsible business practices report, transparency, scaling, and safety sites, as well as the Meta AI Developer Use Guide, and LLaMa Acceptable Use Policy.

On 8 May 2026, Amnesty International sent a letter to Microsoft specifically pertaining to environmental risks, receiving a response on 21 May 2026. In their response, Microsoft reflected that they had engaged an independent firm to conduct a human rights impact assessment of the full lifecycle of the development and deployment of their generative AI products, between September 2024 and April 2025. In addition to outlining their risk monitoring process, Microsoft stated their intention to become “water positive by 2030.”

---

<sup>214</sup> Yaqub Chaudhary and Jonnie Penn, “Beware the intention economy: collection and commodification of intent via large language models”, 30 December 2024, Harvard Data Science Review, Special Issue 5, <https://doi.org/10.1162/99608f92.21e6bbaa>.

<sup>215</sup> Kosmyna et al, 13 November 2025. “Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task”.

<sup>216</sup> Hao-Ping (Hank) Lee, “The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers”, 25 April 2025, Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI ’25), <https://doi.org/10.1145/3706598.3713778>

Microsoft furthermore acknowledged the “increasing global demand for computing power and electricity,” stating that the company is investing in carbon-free energy, and that it intended to “ensure that the electricity cost of serving our datacenters is not passed on to residential customers.” In responding to local community concerns and resistance to their data centres, Microsoft claimed that they are “expanding access to education and digital skills through local partnerships” and “support[ing] sustainability efforts and community-led environmental projects.”

On 9 May 2026, Amnesty International further sent a letter to Intel, who responded on 22 May 2026. Intel noted that “respect for human rights is foundational to Intel and embedded in our Global Human Rights Principles, aligned with the UN Guiding Principles, OECD Guidelines, and ILO conventions. These principles apply across our workforce, operations, and supply chain,” further describing that they maintain human rights due diligence and periodic impact assessments across their enterprise. In relation to general-purpose technologies, Intel stated that they “may not control all end uses of our products; however, we hold suppliers, customers and partners to our standards and enforce compliance with applicable laws.” Nevertheless, the processes mentioned by Intel do not directly impact the findings set out in this publication.

Finally, on 8 May 2026, Amnesty International sent a letter to Amazon. In their response, Amazon states that the company is committed to “respecting internationally recognized human rights, consistent with the UN Guiding Principles on Business and Human Rights,” and describes that its human rights due diligence framework, methodology and tools are deployed across all of its operations, including in “those that support AI workloads.” Amazon noted that they conducted an assessment of Amazon Web Services (AWS) operations and supply chains in 2025, to “to identify and prioritize salient risks in its operations and supply chains and to establish action plans to address them.” Amazon further stated that they “recognize the mineral supply chains underpinning AI hardware, including cobalt and copper, present human rights and environmental risks,” further naming partnerships intended to mitigate risks. Amazon re-stated their objective of becoming “water positive by 2030” and noted working on efficiency and reduction of water and energy use for their data centre operations. In responding to local community concerns and resistance to their data centres in Queretaro, Amazon stated that it had deployed “an air-cooled data center design, which will not require the ongoing use of cooling water in operations,” listing water usage across the year in percentage terms across a number of countries. Amazon did not elaborate substantially on concerns regarding energy usage across data centre operations, not have Amazon’s plans identified in relation to their assessment of AWS been published or clarified to date, to Amnesty International’s knowledge.

No other company has responded to Amnesty International’s findings at the time of publication.



# 6. HUMAN RIGHTS INCOMPATIBILITY OF CURRENT GENERATIVE AI SYSTEMS BASED ON DATA PIPELINE

Through massive data collection without consent, biased training processes and the potential for manipulative outputs, generative AI tools mirror and amplify many of the problems seen in conventional AI systems, such as the machine learning processes undergirding facial recognition for identification (also known as “1:n”) and predictive policing systems, with unavoidable implications for the rights to privacy, equality and non-discrimination, freedom of expression and freedom of thought. This chapter analyses findings related to the design of generative AI tools, explored in Chapter 5, and maps their implications for international human rights standards.

## 6.1 THE RIGHT TO PRIVACY

As set out in section 4.1 above, any interference with the right to privacy must be 1) legally prescribed using clear, precise laws that include adequate safeguards and judicial oversight, 2) legitimate in its aim, and 3) necessary and proportionate to achieve this aim. Additionally, any discriminatory interference is automatically considered unlawful and arbitrary under international law.

Amnesty International has long held that AI-driven surveillance tools that scan, capture and often store data on massive databases curated without individuals’ knowledge and consent, are tools of mass surveillance by design. Amnesty International believes that indiscriminate mass surveillance is never a proportionate interference with the right to privacy. Similarly, Amnesty International’s analysis on the advertising-driven business model of large tech companies such as Google and Facebook, established that such practices create an unprecedented interference with the right to privacy that cannot be compatible with the companies’ responsibility to respect human rights.<sup>217</sup>

Through similar processes, relying on highly resourced web scraping for the collection of massive amounts of data to train generative AI models risk violating the right to privacy. Notably, the bulk and mass collection of training data through web scraping constitutes a mass invasion of privacy by design because such collection includes personal data and is happening without knowledge and consent. Training data has been

---

<sup>217</sup> Amnesty International, *Surveillance Giants* (previously cited), p. 22.

documented to contain personal information from a variety of sources, violating the requirement for legal protection against “arbitrary or unlawful interference” with privacy.

The scale of data collection through tools such as Common Crawl includes indiscriminate collection of personal data, which has been established as being incompatible with the proportionality test. Just as remote biometric recognition technologies depend on mass surveillance by design to curate the reference databases used for processes like facial recognition, which fail the three-part test, most commercially available large generative AI systems (such as LLMs and other multi-modal models) similarly cannot function without web scraping for training data. Furthermore, the generation of synthetic data based on real individuals’ likenesses without knowledge or consent violates the right to privacy.

Generative AI models discussed in this briefing have not, to date, demonstrated restrictions or safeguards that abandon design principles rooted in mass invasions of privacy, nor put in place appropriate safeguards against the capture of personal data. Where generative AI models are not transparent about the provenance of their massive training data and do not include processes by which personal data is excluded in the process of web scraping and from the training data, the practice of web scraping should be considered unlawful. Therefore, generative AI systems built on unlawful web scraping practices fundamentally conflict with the right to privacy as guaranteed under Article 17 of the ICCPR.

## 6.2 THE RIGHT TO EQUALITY AND NON-DISCRIMINATION

Current generative AI systems demonstrate significant risks to the right to equality and non-discrimination, as protected under the ICERD, CEDAW and the ICCPR. Research shows consistent racial, gender and cultural biases in system outputs, which reflect, amplify, reinforce and obfuscate discriminatory patterns in the training data. When used for content moderation, for example, these systems disproportionately affect marginalized communities through biased content removal.

The predominantly English-language training data means that the resourcedness of generative AI models skews towards English. This often takes the form of various biases in the outputs that reflect the social, cultural, linguistic and political norms most present in datasets, reinforcing Western and anglophone cultural and linguistic dominance, while discriminating against and disregarding majority world languages, cultures, ideas and representation.

The generation of synthetic data based on real individuals’ likenesses also violates the right to equality and non-discrimination, especially it perpetuates violence against women and children, with AI-generated CSAM overwhelmingly portraying girls.<sup>218</sup> Children depicted in such material experience significant harm in facing social stigma, isolation, and significant deterioration of their mental health.<sup>219</sup> AI-driven CSAM thus affects children’s right to protection from all forms of violence, abuse and exploitation (Articles 19, 34 UN Convention on the Rights of the Child (CRC) as well as the right to equality and non-discrimination and the right to privacy (Article 16 CRC). In this way, non-consensual image generation constitutes a form of technology-facilitated gender-based violence and violates the rights to live free from gender-based violence, to health, freedom of expression and freedom of peaceful assembly.

## 6.3 RIGHT TO FREEDOM OF EXPRESSION

Generative AI systems pose risks to the right to freedom of expression. Automated content moderation using generative systems can, for example, lead to overbroad censorship, particularly affecting historically marginalized communities. The systems’ inability to properly parse context in different languages and cultures results in suppression of legitimate speech, for example through inaccurate and problematic mistranslations. Additionally, the synthetic content generated on biased training data will limit the range of expression available to some users, as particular racial, ethnic and cultural realities are denied in favour of dominant narratives.

---

<sup>218</sup> Internet Watch Foundation, What has changed in the AI CSAM landscape?, July 2024, p. 23, [https://www.iwf.org.uk/media/opkpmx5q/iwf-ai-csam-report\\_update-public-jul24v11.pdf](https://www.iwf.org.uk/media/opkpmx5q/iwf-ai-csam-report_update-public-jul24v11.pdf) (accessed on 14 May 2026).

<sup>219</sup> For a summary of the impact of real CSAM on children, see Parti/Szabo, *The Legal Challenges of Realistic and AI-Driven Child Sexual Abuse Material: Regulatory and Enforcement Perspectives in Europe*, Laws 2024 (Vol.13), pp. 3-4.

## 6.4 RIGHT TO FREEDOM OF THOUGHT

Generative AI systems also raise new and emergent concerns regarding the protection of the right to freedom of thought. As the case studies in this briefing outline, the manipulation of user intentions and thought processes through predictive suggestions may constitute coercion under Article 18 of the ICCPR, while repeated exposure to synthetic content and algorithmic biases can shape beliefs and mental models without user awareness as a form of manipulation. Crucially, the systems' ability to influence decision-making processes may impair autonomous thought formation and critical discernment, instead facilitating the transfer of biases held within the training data to users of the particular generative AI system in question.

## 6.5 BUSINESSES' HUMAN RIGHTS RESPONSIBILITIES

Under the UN Guiding Principles, companies developing and deploying generative AI systems have generally fallen short of meeting their responsibilities to:

- Carry out adequate human rights due diligence during the development and deployment stages of systems;
- Provide a bare minimum of transparency about data collection and processing practices;
- Take measures to prevent and mitigate adverse human rights impacts that could stem from their systems.

Generative AI developers have generally relied on having “usage policies” to varying degrees, which ask users to commit to using their products responsibly. Without enforceability, these policies do not go far enough in ensuring human rights-compliant uses of generative AI products, instead shifting the onus from company to user, without taking responsibility for design, development, deployment and policy choices that could mitigate downstream human rights abuses.

This is complicated by the question of whether a particular generative AI system is “open” or “closed.” While some commercially available generative AI products are proprietary, many generative AI developers increasingly make their products available as open-source tools. While open-source availability has been hailed as providing greater transparency into the architecture of these systems and for widening access to repositories of sophisticated computing capabilities, it rarely fulfils the need for accountability. It furthermore significantly complicates the issue of where accountability and liability rest. The open nature of some generative AI tools does not itself solve human rights issues across the supply chain and lifecycle of the product and can even be used as a means of absolving developers of liability and accountability for design features that involve human rights harms.

While it is crucial that systems can be independently and transparently assessed, investigated and understood, developers have a responsibility to ensure that they are actively preventing abuses of international human rights standards that could result from the product's use, regardless of whether the system in question is closed- or open-source.

# 7. CONCLUSIONS AND RECOMMENDATIONS

## 7.1 CONCLUSION

This briefing demonstrates that many of the fundamental design features of mainstream standalone generative AI systems appear incompatible with aspects of international human rights law and standards, particularly stemming from unlawful web scraping baked into the data pipeline feeding these tools. This is further compounded by the increasing human rights risks associated with how these systems are deployed and used. Where the design of their products depends on mass data collection practices that violate privacy rights, and where products generate outputs that systematically discriminate against marginalized groups, these present significant incompatibility with the right to privacy and the right to equality and non-discrimination.

Without significant changes to how these systems are developed and deployed, they will continue to undermine international human rights laws and standards. Whether proprietary or open source, these systems require urgent attention to establish clear frameworks for accountability and human rights protection before their impacts become more deeply entrenched in society. Amnesty International therefore calls for a prohibition of standalone generative AI systems identified to have been built on unlawful web scraping, as well as those which have demonstrated patterns of discrimination.

## 7.2 RECOMMENDATIONS

### TO STATES

- Implement a prohibition on standalone generative AI systems that have been built using unlawful web scraping, defined as the bulk and mass collection of training data through the World Wide Web, without protection against non-consensual collection of personal data.
- Enact legislation requiring transparency regarding training data collection practices and accountability across AI supply chains, and further:
  - Require in law that technology companies, including those developing and deploying generative AI systems, carry out ongoing and proactive human rights due diligence to identify and address human rights risks and impacts related to their global operations. This must include clear regulatory frameworks requiring mandatory human rights impact assessments before the deployment of generative AI systems.
  - Establish mandatory environmental impact assessments for data centres and computing infrastructure supporting generative AI systems, with clear reporting requirements.
  - Ensure human rights impact assessments are required for the entire supply chain associated with the deployment of an AI product, including with regards to the production of GPUs,

chips and minerals required to build them, which may be situated in adverse human rights contexts.

- Ensure meaningful consultation by independent bodies with affected communities, particularly those historically marginalized or discriminated against, throughout the lifecycle of the product.
- Where AI deployments are identified as exacerbating existing inequalities or creating new forms of discrimination, to cease their use.
- In all development, deployment and use of any AI system, guarantee access to effective remedy for human rights abuses linked to the impacts of technology companies, wherever the harms occur, including harms resulting from the operations of their subsidiaries, whether foreign or domestic. Redress mechanisms should be made easily accessible and understandable to enable individuals to file complaints when their rights have been infringed.
- Invest in, encourage and promote the implementation of effective digital educational programmes to ensure that individuals understand their rights, including their right to seek an effective remedy against any data protection, privacy or other human rights abuse, when accessing digital services.

## TO COMPANIES

- Cease the practice of unlawful non-consensual web scraping of personal data for AI training purposes.
- Conduct comprehensive human rights due diligence throughout the development and deployment lifecycle of generative AI systems, including publicly reporting on any risks identified and the risk mitigation measures taken to address them.
- As part of due diligence, provide full transparency (or publicly disclose) of the risks and abuses identified in relation to data collection and processing practices, supply chain impacts, including the environmental footprint of operations, as well as the risk mitigation measures taken to address any risks and abuses.
- Establish clear accountability mechanisms and grievance procedures for individuals and communities affected by their systems, which are accessible to all, sufficiently clear, responsive and timely.
- Take immediate action to bring generative AI products in line with the right to equality and non-discrimination by proactively identifying and eliminating discriminatory biases in their systems. Where this is not possible, models should be discontinued from commercial and public usage.
- Ensure meaningful and effective consultation with affected communities in relation to the design and development of lawful AI models and their impacts, particularly those communities that have been historically marginalized or discriminated against.
- Ensure that content moderation guidelines, rules and practices are based on – and consistent with – international human rights law and standards and implemented on the basis of equality and non-discrimination.
- Ensure appropriate investment in local-language resourcing in content moderation throughout the world, with a particular emphasis on resolving existing inequalities that disproportionately impact Global Majority countries.
- Human rights impact assessments should be published on a regular basis and should include detailed information on risks and mitigating measures taken with respect to specific countries (especially where systems may have a greater impact due to political conflicts or humanitarian emergencies), specific categories of users such as children and young people, and specific product changes.

## TO UN BODIES AND REGIONAL ACTORS

- Develop international standards for protection against the development and deployment of generative AI systems that are incompatible with international human rights law by design, in line with UNGA resolution A/78/L.49 “Seizing the opportunities of safe, secure and trustworthy artificial intelligence

systems for sustainable development” and clause 20(b) of UNGA resolution A/RES/78/213 “Promotion and protection of human rights in the context of digital technologies”.

- Establish clear frameworks for accountability and redress when rights violations occur.
- Support independent research and monitoring of generative AI systems’ human rights impacts, including across the mandates of special procedures.

**AMNESTY INTERNATIONAL  
IS A GLOBAL MOVEMENT  
FOR HUMAN RIGHTS.  
WHEN INJUSTICE HAPPENS  
TO ONE PERSON, IT  
MATTERS TO US ALL.**

## CONTACT US



[contactus@amnesty.org](mailto:contactus@amnesty.org)



+44 (0)20 7413 5500

## JOIN THE CONVERSATION



[www.facebook.com/amnesty](https://www.facebook.com/amnesty)



@Amnesty

# UNLAWFUL BY DESIGN

## EXPOSING THE HUMAN RIGHTS COSTS OF GENERATIVE AI

This briefing examines how standalone generative AI systems, based on unlawful web scraping, are in conflict with international human rights law (IHRL) and standards through their design, development and deployment. While these technologies promise sophisticated automation and efficiency, they rely on data collection and model training practices that abuse privacy rights, enable discrimination, and threaten freedom of expression and thought.

The briefing looks at the models powering some of the most popular publicly available standalone generative AI tools, including ChatGPT, Dall-E, Gemini, Midjourney, Stable Diffusion, and DeepSeek. The models under analysis include large language models (LLMs), in particular generative pre-trained transformers (GPT), generative adversarial networks (GANs), variational autoencoders (VAEs) and diffusion models (DM).

Amnesty International finds that standalone generative AI systems, based on unlawful web scraping, depend on mass invasions of privacy by design, and are fundamentally incompatible with IHRL. As such, Amnesty International is calling for a prohibition of such systems.