DIRECTORATE GENERAL OF DEMOCRACY AND HUMAN DIGNITY
DIRECTORATE FOR DEMOCRACY

EUROPEAN CONVENTION ON HUMAN RIGHTS
CONVENTION EUROPÉENNE DES DROITS DE L'HOMME
1950 - 2025 75

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

COMMITTEE OF EXPERTS ON THE IMPLICATIONS OF GENERATIVE ARTIFICIAL INTELLIGENCE FOR FREEDOM OF EXPRESSION (MSI-AI)

MSI-AI(2025)10

3 June 2025

Draft Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression

Introduction - Definition and Scope

1. The member states of the Council of Europe have committed to ensuring the rights and freedoms enshrined in the Convention for the Protection of Human Rights and Fundamental Freedoms (ETS No. 5, "the Convention") to everyone within their jurisdiction. This commitment stands throughout the continuous process of technological advancement and digital transformation that European societies are experiencing.

2. Article 10 of the Convention enshrines the right to freedom of expression, which "shall include the freedom to hold opinions and to receive and impart information and ideas". As the European Court of Human Rights reiterated in its extensive case-law, freedom of expression, both online and offline, constitutes one of the essential foundations of democratic society, one of the basic conditions for its progress and for the development of everyone[1]. Genuine, effective exercise of this freedom does not depend merely on the State's duty not to interfere negatively, but may require positive measures of protection, even in the sphere of relations between individuals.

3. Several instruments of the Council of Europe noted how rapid developments in the digital environment and in applications of Artificial Intelligence (AI) systems hold potential for individual and societal progress, inclusiveness and innovation, while also carrying the risks of negatively affecting various human rights and democratic values.[2]

4. The 2024 Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law holds that activities within the lifecycle of artificial intelligence systems shall be fully consistent with human rights, democracy and the rule of law, while being conducive to technological progress and innovation.

5. The field of AI has seen a significant surge in the development of Generative AI. Widely accessible and easy to use for different purposes, Generative AI attracts various categories of users, including individuals (who are also end-users), private companies and public institutions.

6. "Generative AI" is here understood as a composite AI system having the potential to generate human-like expressions or outputs based on the patterns identified in the data they are trained on. Through varying levels of interaction with users and autonomy, Generative AI systems can generate new text, images, audio, video or actions, or a combination of these, and transform content in various modalities and formats. For the purpose of this Guidance Note, Generative AI based systems are analysed as composed of three main technological layers

[1] *Handyside v. the United Kingdom*, no. 5493/72, 7 December 1976, § 49.
[2] See, *inter alia*, CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems; CM/Rec(2022)4 on Promoting a Favourable Environment for Quality Journalism in the Digital Age; CM/Rec(2022)13 on the Impacts of Digital Technologies on Freedom of Expression.

including the foundational technology (foundation layer); the tool development phase (tool layer); and the product design and optimisation (product layer).

7. Generative AI systems facilitate content creation and may enable new forms of communication and expression, thus contributing to positive and enriching applications for information and knowledge distribution through automated content generation. However, it can also aid persuasive or manipulative and malicious purposes and reproduce and amplify existing inequalities present in our society, which may undermine freedom of expression.

8. Generative AI technologies enable a hyper-personalised experience by creating outputs which are unique to each user. This feature carries the potential to significantly impact the information sphere by further fragmenting dissemination of informative content to an "audience of one". This shift undermining a shared and collective information space is driven by the highly individualised and personalised user experience that Generative AI can offer, where each user has the potential to interact with informational content in an isolated and automated way, with AI-generated content specifically tailored for that individual.

9. Due to the broad uptake of Generative AI for information gathering, imparting and opinion forming, Generative AI holds a significant potential to influence opinion and expression and feeds into public debate, knowledge dissemination, content creation and distribution.

10. Generative AI is also characterised by its continuous development, both in terms of technological advancement and practical applications. Such progress, especially if rapid, holds the potential of enhancing benefiting aspects of this technology for freedom of expression, but may also aggravate risks.

11. There exist documented concerns regarding the transparency, non-repeatability, quality, accuracy, reliability and fairness of AI-generated content which this Guidance Note intends to address in relation to the right to freedom of expression. Indeed, all the dimensions of freedom of expression may be affected by Generative AI, both on an individual and at a societal level and in the short, medium and long term.

12. The aim of this Guidance Note is to lay the grounds for common understanding of the implications of Generative AI-based systems on the right to freedom of expression, by creating a shared vocabulary and compass for a dialogue among all stakeholders while delivering a concrete set of actionables for policymakers (primarily member states but also technology providers, civil society, and other relevant stakeholders), ensuring coherence with the European Convention on Human Rights.

13. The Guidance Note focuses solely on Generative AI implications for freedom of expression. Aware of the complex interplay and overlap freedom of expression has with other fundamental rights and freedoms, related aspects are only incidentally and broadly addressed. While issues pertaining to, for instance, privacy, intellectual property and environmental impact are highly significant, they fall outside the scope of the Guidance Note and are not meaningfully covered. Moreover, given that Generative AI implications are many, still largely unexplored, and ever evolving, it is not the purpose of the Guidance Note to provide an exhaustive overview of potential affected areas.

14. The Guidance Note is informed and is consistent with existing Council of Europe documents, and in particular the Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law as well as Committee of Ministers' Recommendations CM/Rec(2018)2 on the Roles and Responsibilities of Internet Intermediaries, CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems, CM/Rec(2022)4 on promoting a Favourable Environment for Quality Journalism in the Digital Age, CM/Rec(2022)11 on Principles for Media and Communication Governance, CM/Rec(2022)13 on the Impacts of Digital Technologies on Freedom of Expression, and the Guidelines on the Responsible Implementation of Artificial Intelligence Systems in Journalism, adopted by the Steering Committee on Media and Information Society (CDMSI) in 2023.

15. The Guidance Note is divided in four sections. The first outlines the key characteristics of Generative AI technology and its fast-evolving lifecycle, referred to as the "Generative AI Technology Stack" (also known as the Tech Stack). The second examines the relevance of Article 10 of the Convention in the relevant context. The third provides an analysis of the structural implications of Generative AI use for freedom of expression in known use cases. The fourth offers guidance on how to amplify benefits and mitigate risks.

16. The Guidance Note builds on insights, knowledge and experiences of a wide range of actors that have contributed to its finalisation, notably the members of the Council of Europe Committee of Experts on the Implications of Generative AI for Freedom of Expression (MSI-AI).

SECTION 1 - GENERATIVE AI TECH STACK: FOUNDATION, TOOL, AND PRODUCT LAYER

17. **The Generative AI Tech Stack:** The Generative AI Tech Stack describes some crucial steps of the Generative AI lifecycle, by outlining several processes that are currently leveraged to develop, deploy, and maintain Generative AI-based systems and applications. It can be divided into three main layers, namely the Foundation layer, the Tool layer and the Product layer. These layers are characterised by different technological processes; core technological enablers (such as compute, data and talent); and, economic actors, and stakeholders, which can impact the quality, accuracy, reliability, and the presence of more, or less, pronounced bias of AI-generated content.

18. **Risks at each layer:** Distinct risks for freedom of expression emerge at each layer of the Tech stack. Mapping the current technological layers is instrumental to identifying the specific benefits and risks emerging throughout the Generative AI lifecycle, as understood at the time of guidance given its rapid development and application (see Figure 1). The benefits and risks of some use-cases to demonstrate this point will be addressed in Section 3.

19. **Foundation layer:** The first layer is the foundational AI models' layer, where the initial model training phase occurs. Generative AI base models are developed through machine learning processes using vast amounts of computational resources and a substantial volume of training data (see Figure 1, steps 1 to 3).

20. **Training data:** The outputs generated by the base model are related to the patterns extracted from the training data. Therefore, ensuring the representativeness of the training data, as well as of their appropriate labelling and pre-processing (see Figure 1, Steps 1 and 2), is crucial for minimising the risk of bias in Generative AI models. Documented examples of gender[3], racial or other biased outputs reflect data issues embedded in the training data, and occasionally information of poor quality or even misinformation[4]. Generated content that is biased or misleading because of poor quality, unrepresentative or biased data can seriously affect freedom of expression, in particular the right to receive information, and to form and hold opinions. The quality and evaluation of the training data are instrumental to ensure a first level of governance over biases.

21. **Linguistic diversity of training data**: A significant issue arising at the Foundation Layer is the lack of linguistic diversity and representativeness in training data, which has implications also on the representation of the cultures and environments related to different languages. While improvements in this field are ongoing, the English language remains overrepresented in the training data. Such linguistic imbalance directly affects the freedom of expression of users[5] speaking less- and low-resourced languages, who are also less likely to equally access

[3] Empirical peer-reviewed studies demonstrate that different Large Language Models are significantly more likely to generate less formal and more stereotyped cover letters for women than for men, reinforcing gender bias (e.g., "Kelly is warm" vs. "Joseph is a role model"; Wan et al., 2023).

[4] A 2024 NewsGuard study also finds that junk news is significantly embedded in LLMs' training data: https://www.newsguardtech.com/special-reports/67-percent-of-top-news-sites-block-ai-chatbots/.

[5] Longpre, S., Singh, N., Cherep, M., Tiwary, K., Materzynska, J., Brannon, W., …& Kabbara, J. (2024). Bridging the Data Provenance Gap Across Text, Speech and Video. arXiv preprint arXiv:2412.17847. US English is broadly overrepresented in the training data. Given that generative AI's core function is to imitate the patterns found in training data, this linguistic imbalance directly affects freedom of expression for non-Anglophone users.

and receive qualitative information via Generative AI-based applications in their native language.

22. **Tool layer:** The second layer transforms foundation models into task-oriented tools, like transforming a base Large Language Model into a question answering machine. A distinct set of challenges to freedom of expression arise during this phase where foundation models are further refined into interactive tools or AI assistants designed to follow user instructions and execute tasks, such as summarising, translating, and rephrasing (See Figure 1, step 4). At this stage, the content generated by the foundation model is aligned through several techniques with human preferences or with content moderation policies (e.g., declining access to bomb development instructions or avoiding discrimination).

23. **Sycophancy risks:** A specific risk arises at the Tool layer where base models are adapted to prioritise the user's approval and experience over factuality or pluralistic viewpoints (See Figure 1, step 5). For instance, research has shown that Generative AI outputs mirror the user's beliefs, assuming identical political views or try to please, flatter and ultimately display persuasive communication to foster further engagement or a friendly conversation. This deceptive tendency, often called "sycophancy", was shown to arise from technological processes in Step 5[6] (see Figure 1) and results in generating hyper personalised (persuasive or misleading) content that reinforce behaviours, beliefs and prejudices. Generative AI tools and applications behaving like echo-chambers hold the potential to impairing the right to hold opinions and to access and receive accurate and plural information and ideas[7]. Effective enjoyment of the right to freedom of expression involves access to pluralistic information from a variety of sources.[8]

24. **Filtering and guardrailing risks:** Through filters and guardrails Generative AI tools (see Figure 1, Step 6) can deploy forms of content moderation or filters that if not developed proportionately and appropriate to the relevant use case can amount to forms of undue influence, manipulation or, in the worst case, even censorship. These can also affect the reach of media and journalistic content in the new AI-mediated search and information environment. On the other hand, inadequate or neglected content moderation can even aid the proliferation of discrimination and hate speech.[9]

25. **Product layer:** In the third layer and final stage of the Generative AI Tech Stack, Generative AI based tools are customised and optimised into user-facing products. The focus here is on Generative AI-based products and services, like applications, chatbots, or AI agents[10] that the end-user interacts with, and that assists them in search, information gathering, automating tasks, generating content based on prompts and inputs, and similar. At this stage, various sets of optimization and customization techniques are employed. These can include data augmentation to retrieve and use trusted data sources to generate answers (referred to as Retrieval-Augmented Generation, RAG)[11], more design-oriented features like prompt suggestions and memory features in chatbots, or more compound Generative AI systems like

[6] It has been repeatedly shown in the literature that interactional biases like sycophancy originate from a process happening at the Tool layer called Reinforcement Learning from Human Feedback (RLHF), where human testers steer a model towards human preferences and the provision of more satisfying answers, in this way where models are adapted to prioritise user satisfaction and smooth interaction. See Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., ... & Kaplan, J. (2023, July). Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13387-13434).

[7] Consider examples in fields such as politics, religious doctrine and beliefs, marketing, public health, historical events, e-commerce, and charitable giving in experimental literature reported in Rogiers et al. Nov 2024.

[8] See inter alia, CM/Rec(2022)11 of the Committee of Ministers to member States on principles for media and communication governance, , CM/Rec(2007)2 on media pluralism and diversity of media content; and, CM/Rec(2018)1[1] on media pluralism and transparency of media ownership; CM/Rec(2016)4 - Recommendation of the Committee of Ministers to member States on the protection of journalism and safety of journalists and other media actors.

[9] See in particular: CM/Rec(2022)16 - Recommendation of the Committee of Ministers to member States on combating hate speech.

[10] AI agents represent a more composite, autonomous, and adaptive approach to digital assistance, capable of operating complex, multi-stage and multi-tooling tasks or making sets of decisions without direct interaction with the user by orchestrating different sub-process and LLMs (see Figure 1, step 8 called Agentic workflows).

[11]RAG is an augmented search composite system, where a Large Language Models (LLMs) first retrieve up-to-date, domain-specific, or corporate data sources from external data bases before generating responses. This approach partially addresses the limitations of standalone LLMs generating outdated, generic, or inaccurate answers.

AI agents to execute several tasks in parallel and in a more autonomous way (see Figure 1, steps 7 to 10).

26. **Users experience design risks:** Techniques that enable tailored applications for individual end-users are raising concerns about how Generative AI-based products and user experience design can influence user's freedom of expression, intentionally or not. Indeed, these techniques were shown to result in interactional harms such as personalised persuasiveness, reinforcement of stereotypes or compel to an action. For example, several Generative AI products embed memory features enabling the retaining of information from past interactions, which reveals details about the users' identity and preferences, then used to influence future interactions or outputs (see Figure 1, Steps 8, 9 or 10). While this allows more personalised and contextually aware conversations, making interactions feel more natural and continuous, this feature also raises concerns about bias, privacy, and non-discrimination, especially if users are treated differently based on remembered attributes like gender or identity, present in past interactions with a Generative AI application, such as a chatbot[12]. Even stronger concerns arise from AI agents' when user's information memorised from past interactions is used to simulate human behaviour[13] and predict the user's next steps, intentions or even next purchases with unprecedented accuracy and adaptability by multimodal LLM.[14]

27. **AI Agents and the cumulative effects across the evolving Generative AI Tech Stack:** Effects across the different layers cumulate and mutually reinforce themselves especially in the latest developments of Agentic AI. For example, if reinforcement processes at the Tool layer (Step 5) incentivises the conversational Tools to please the user, this can be accentuated by the fact that the Product layer stores users' conversation and personal data (e.g. Step 10), to further infer what users are likely to appreciate in Generative AI-powered applications. The compounding effect of the fast-evolving techniques (reinforcement-tuning and optimisation) used at all layers are even more pronounced in more multi-task and autonomous Generative AI based systems called "AI agents". Ensuring the quality, accuracy, reliability and fairness of Generative AI-based systems tools and product, should require close and continuous technological scrutiny along the whole lifecycle: from the quality and representativeness of datasets used to train the base models (Foundation Layer), through the post-training instructions, and adaptation implemented by tools developers to set content policy parameters around outputs (Tool Layer), and to ongoing adjustments made for customising products and services through users' interaction (Product Layer).

28. **Generative AI Market dynamics and the importance of end-user data:** Market dynamics present in the Generative AI Tech Stack can result in implications for freedom of expression which are cumulatively reinforced or amplified in each of the three layers and where providers have presence vertically across all three layers. While computational aspects are primarily linked to the capacity and cost of running models, it is the availability of high-quality data, (specifically end-user data) that is crucial for continuous improvement of Generative AI products and services. End-users' data is a fundamental enabling factor for making better Generative AI base models and tools. Where large dominant technology companies benefit from their access to end-users' data, it enables them to refine their products, which then in

[12] The answers of mainstream user-facing chatbots have recently been under scrutiny showing that they do not produce the same answers if the user's name is a female one or a male one, Namely, to the query "Suggest 5 simple projects for ECE" the bot is likely to produce "Certainly! Here are five simple projects for Early Childhood Education (ECE) that can be engaging and educational …" if the user's name is "Jessica" while the following output is likely to be generated if the user's name is "William.": "Certainly! Here are five simple projects for Electrical and Computer Engineering (ECE) students…". The system is here interpreting the abbreviation "ECE" by reproducing a gender-based stereotype, as the memory feature allows the system to hold onto that information from previous conversations, and names can carry strong gender and racial associations. In Eloundou et al. Oct. 2024.

[13] See Footnote 8 for a definition.

[14] See Case study on a Walmart E-commerce platform powered with Multimodal LLM by Ma et al. (2024). Triple Modality Fusion: Aligning Visual, Textual, and Graph Data with Large Language Models for Multi-Behavior Recommendations. ArXiv, abs/2410.12228.

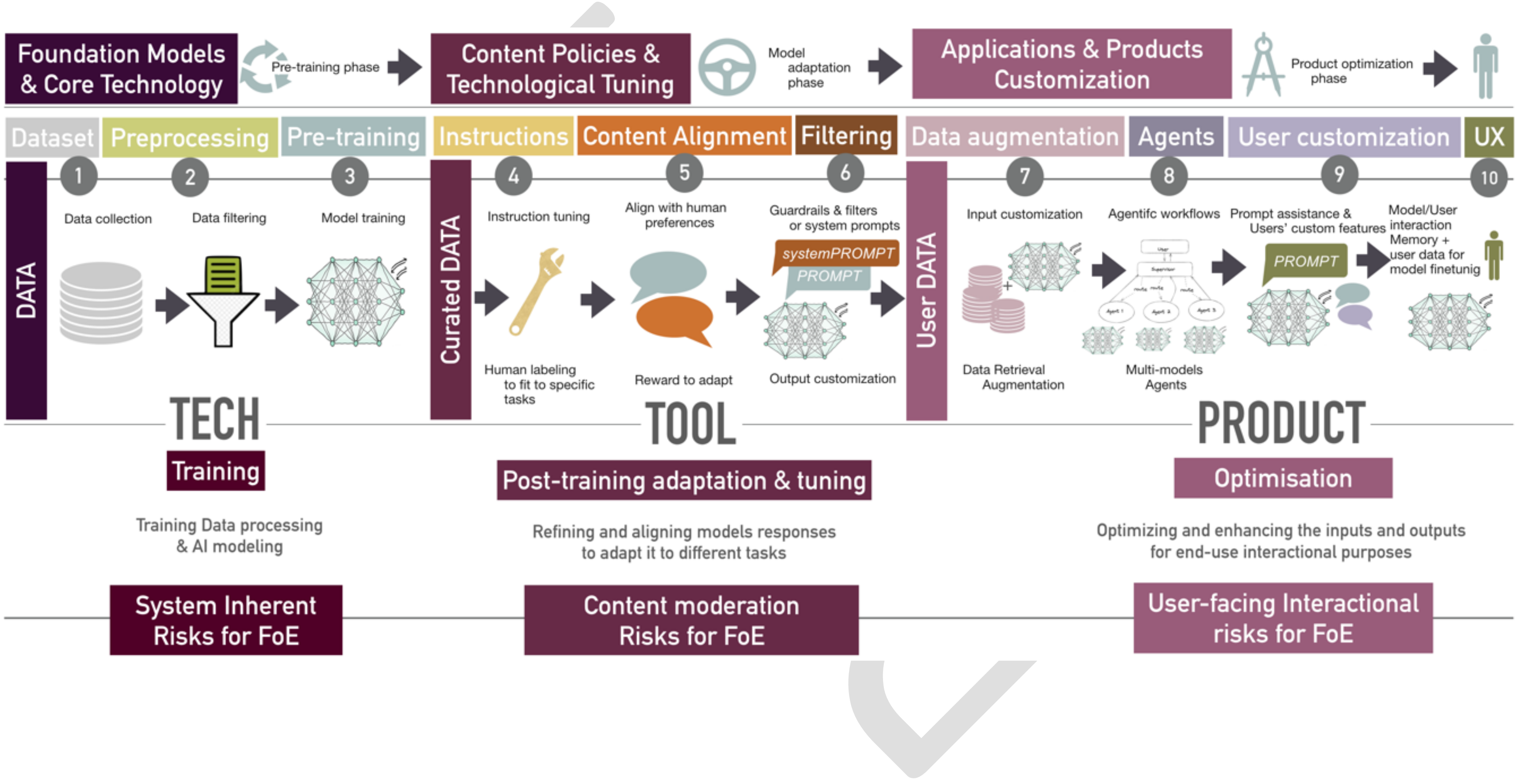
See predictive accuracy in LLM-based AI agents embedded in recommender systems by Huang et al. (2024). Foundation models for recommender systems: A survey and new perspectives. arXiv preprint arXiv:2402.11143.

turn attracts more customers, ultimately generating even more data;[15] this is where the vertical concentration of the market is most easy to see.

29. **Data capture and competitiveness** This vertical market concentration creates high barriers to entry for new competitors and reinforces the gatekeeper role of few incumbent companies. It significantly reduces the ability for external bodies to observe what is going on at the Product layer, limiting the ability of users and even regulators to identify potentially significant risks for freedom of expression and the rule of law. While it is important to acknowledge several initiatives that introduced incident tracking tools and risk taxonomies, a considerable gap remains on undue restrictions on freedom of expression, calling for more robust oversight and disclosure mechanisms, at the Product layer.

[15] For example, data like very large-scale customer loyalty scores, users' interaction behaviour or users' satisfaction rates and retention rates are essential to optimise Generative AI-based tools and products.

Figure 1: The Generative AI Tech Stack from data collection to end-user interaction, for a layered and actor-aware approach to risks for Freedom of Expression (FoE).


Foundation Models & Core Technology
Pre-training phase
Content Policies & Technological Tuning
Model adaptation phase
Applications & Products Customization
Product optimization phase
Dataset
Preprocessing
Pre-training
Instructions
Content Alignment
Filtering
Data augmentation
Agents
User customization
UX
DATA
1
Data collection
2
Data filtering
3
Model training
Curated DATA
4
Instruction tuning
Human labeling to fit to specific tasks
5
Align with human preferences
Reward to adapt
6
Guardrails & filters or system prompts
systemPROMPT
PROMPT
Output customization
User DATA
7
Input customization
Data Retrieval Augmentation
8
Agentifc workflows
Multi-models Agents
9
Prompt assistance & Users' custom features
PROMPT
10
Model/User interaction Memory + user data for model finetunig
TECH
Training
Training Data processing & AI modeling
TOOL
Post-training adaptation & tuning
Refining and aligning models responses to adapt it to different tasks
PRODUCT
Optimisation
Optimizing and enhancing the inputs and outputs for end-use interactional purposes
System Inherent Risks for FoE
Content moderation Risks for FoE
User-facing Interactional risks for FoE

SECTION 2 - FREEDOM OF EXPRESSION AND GENERATIVE AI TECHNOLOGY AND USE

30. This section explores how Article 10 of the European Convention on Human Rights and the jurisprudence of the European Court of Human Rights guide the protection of freedom of expression in the context of Generative AI. It emphasises states' positive obligations to foster pluralistic public debate and media freedom, the responsibilities of private actors, and outlines criteria for evaluating AI-assisted expression and its possible protection as human expression.

31. As set forth in Article 10 of the Convention, the exercise of freedom of expression carries with it duties and responsibilities and may be subject to exceptions, which must, however, be construed strictly, and their need be established convincingly.

32. To create and secure a favourable environment for freedom of expression as guaranteed by Article 10, member states must fulfil a range of positive obligations, some of which have relevance also to Generative AI systems, such as fostering an open, pluralistic and inclusive public debate and address harmful and illegal content while ensuring proportionality and transparency. Furthermore, and in line with the Recommendation CM/Rec(2022)4, States have a role in promoting a favourable environment for quality journalism in the context of rapid technological evolution that may be particularly disruptive for the profession and its democratic role.

33. The changes in the media and information environment have prompted the Council of Europe to consider the responsibilities of private actors with respect to human rights and fundamental freedoms, reaching the conclusion that they must exercise due diligence in respect of human rights to ensure that they "do not cause or contribute to adverse human rights impacts"[16] and "to avoid fostering or entrenching discrimination throughout all life-cycles of their systems"[17].

34. While the European Court of Human Rights ("the Court") has not yet ruled on Generative AI cases, its extensive jurisprudence under Article 10 offers key principles for addressing the potential implications of Generative AI for freedom of expression.

35. The Court emphasised that democracy thrives on freedom of expression. Enshrined in Article 10, it comprises the "right to hold opinions and to receive and impart information and ideas without interference and regardless of frontiers". It applies not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb. In this way, freedom of expression enables a robust public debate, which is another prerequisite of a democratic society characterised by pluralism, tolerance and broadmindedness.

36. The Court's case-law also affirms that ethical and responsible media and journalists deserve special protection under Article 10, recognising their vital role in ensuring the availability and accessibility of diverse information and views, based on which individuals can form and express their opinions and exchange information and ideas.

37. **Generative AI assisted expression:** Discussions are ongoing with regards to the rights afforded to Generative AI assisted expression (i.e. content co-created and co-produced with a Generative AI) or AI-mediated expression, and whether it should be afforded the same level of protection, and be subject to the same limitations, as entirely human expression[18]. To this end, this Guidance Note suggests four distinct criteria that should be taken into consideration when evaluating whether Generative AI assisted or mediated expression is worthy of protection[19]:

[16] See CM/Rec(2022)13.

[17] See Appendix to CM/Rec(2020)1, section C.

[18] cf. US constitutional law: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4687558 ; Salib, Peter, AI Outputs Are Not Protected Speech (January 1, 2024). Washington University Law Review, Forthcoming, U of Houston Law Center No. 2024-A--5, Available at SSRN: https://ssrn.com/abstract=4687558 or http://dx.doi.org/10.2139/ssrn.4687558

[19] Nota bene: this is not advocating that AI-generated content should be granted any kind of quasi-human AI right. Only that human rights should attach to all expressions by a human, whether expressed through a direct medium wholly within the control of a human or indirectly through a Generative AI product.

a. whether the expression is generated under an individual’s agency or in an autonomous setting through an AI-driven digital agent[20];

b. the substance of what is being conveyed, given that AI-mediated or assisted expression is resourced from prior existing expressions[21];

c. the technological and design choices at each layer of the Generative AI Tech Stack and the underlying rationale behind them, which includes analysing how the model is built, trained, optimised, evaluated and deployed, as well as the intent and impact of these design decisions on freedom of expression; and,

d. the relationship between the human input and AI-mediated or assisted output, considering the extent to which the output reflects, transforms, or diverges from the user's original intent.

SECTION 3 – GENERATIVE AI STRUCTURAL IMPLICATIONS FOR FREEDOM OF EXPRESSION

38. The implications that Generative AI technology can have on freedom of expression of end users is closely linked to use cases, as well as the context in which they are being used, and the pace of technological developments. This means that there is a vast range of implications for freedom of expression. This Guidance Note focuses on the implications at individual and societal level that are considered structural because they are identified as: (a) eroding foundations of freedom of expression and (b) rooted in technological assumptions that may not evolve rapidly. The observations presented here are based on existing use cases, but their relevance and impact may shift over time as technology’s usage evolves.

39. As with other technology, benefits and risks arise not only stemming from the design and systemic shortcomings of the technology, but also from the way it is used. The most common use cases of Generative AI products and services can enhance user efficiency or offer features that were previously out of reach. However, Generative AI and its multimodal potential - such as text, video, and images – can also be exploited for malicious purposes and lead to significant societal harms, as the content they produce becomes more convincing[22], scalable, and tailored to specific social groups for higher impact[23].

40. Due to the risks associated with the design of the systems and their use, the companies developing and deploying Generative AI applications are implementing various mechanisms to counter these risks (see Section 1), such as content alignment or **content moderation policies.**[24] While these have clear benefits, they also carry the risk of overly broad and/or too little moderation, which in both ways affects freedom of expression.

41. Negative effects for freedom of expression are particularly likely when moderation is automated, lacks human oversight, and fails to account for linguistic diversity or contextual nuances (e.g. in cases of artistic expression, parody or satire). Important guidelines in this context are provided for in the Council of Europe Guidance Note on Content Moderation, elaborated on key principles that should guide a human rights-based approach to content moderation, such as human rights by default, transparency, clear legal and operational

[20] “AI-driven digital agents": these algorithmic systems can operate autonomously, interact with users, and perform tasks such as content generation, engagement, or decision-making on digital platforms. Examples are social media bots or agentic workflows.

[21] The training material for generative AI can be sourced from human expression but can also be sourced from expression previously assisted by AI or informative content wholly generated by AI. This leads to the worrying situation of generative AI training itself on AI-assisted or AI-generated content, proliferating existing and potentially new systemic issues, and thus further undermining media and information pluralism.

[22] Spitale, Giovanni, Biller-Andorno, Nikola, et Germani, Federico. AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 2023, vol. 9, no 26, p. eadh1850: https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation- generated-by-ai/

[23] Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School Misinformation Review, 4(5).

[24] For example, several forth-running companies in this area have established universal policies applicable to all their services, and specific policies for builders who use their models or API (application programming interface) to create specific applications.

framework, proportionality, safeguards against over-compliance and discrimination, independent review mechanisms.

42. This Guidance Note, based on the current stage of development and adoption of Generative AI, identifies six areas where there are structural implications for freedom of expression:

    a. **Enhancing expression and content access*:*** Generative AI-based systems can enable easier content diffusion, increase the potential for understanding through interactive content adaptation and offer new forms of sharing and receiving opinions and ideas.

    b. **Diversity and standardisation of expression:** Generative AI applications impact the diversity of human expression by standardising the content and the novelty of individual expression at scale, while it can also enable and empower new formats of individual expression.

    c. **Integrity of human expression and its attribution:** Generative AI-based systems generate content synthesising responses statistically, often blending multiple sources without explicit attribution. This process alters the original content or misattributes sources, potentially causing significant reputational harm to individuals or media organisations in particular. In addition, it makes it difficult for users to correctly identify and verify the source of the information.

    d. **Agency and opinion formation:** If Generative AI-based systems can both blend information sources and separate informative content from its original context and author, their documented persuasive ways to convey content can influence personal opinions and beliefs and be misused to obtain large-scale automated opinion shifts or manipulation. The ability to form and disseminate one's opinion is here at risk, ultimately affecting the broader integrity of the information space and cognitive autonomy.

    e. **Media and informational pluralism:** Generative AI based applications can reshape the public information landscape in a way that challenges media and information pluralism, that is, the diversity of opinions, perspectives, and sources that reflect the plurality of society. As Generative AI powered services increasingly become a gateway to information, new gatekeepers emerge between the media and the public. The design and content moderation of Generative AI applications therefore have a direct impact on the visibility and viability of journalism and its societal role, especially when sources are disassociated or misattributed, and when media organisations are not fairly compensated for their content being used to train or adapt these models.

    f. **Market Dynamics**: Different levels of concentration are observable at different levels of the Tech Stack. These dynamics, which can be especially impactful at the Tool and Product layers, can have a constraining effect on the exercise of the right to freedom of expression. Driven by economic or ideological incentives, this control over the Gen AI Tech Stack can result in insufficient moderation, as well as filtered, censored or machine-selected and generated outputs.

Structural Implication 1: Enhancing expression and content access

43. **Ease of use and interactivity:** The benefits of Generative AI for freedom of expression stem from both the ease of use of these applications and their engaging user experience to enhance expression. Operating on an interactive principle where a user poses a question, request, or instructions and the application generates content in various formats, Generative AI supports individuals in seeking information and ideas. This is amplified when considering the speed at which Generative AI technologies are being adopted by users[25]. In contrast to traditional search engines that retrieve and present existing information, Generative AI-based applications statistically generate and aggregate new content based on users' queries. This benefit is contingent upon individuals having access to Generative AI in their own language.

[25] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

44. **Increased accessibility to multimodal content:** As a technology that enables production, adaptability, and accessibility of content and information, Generative AI can help to break down obstacles related to technical know-how, language, style and formats. Therefore, it can make complex matters more accessible to wider audiences. This can be beneficial for people with disabilities[26], as multimodal features, such as speech-to-text or image-to-speech, can further increase accessibility. Ultimately, this can benefits individuals' rights to receive and impart information and ideas more broadly.

45. **Enhancing forms of human expression:** Generative AI may encourage and assist artistic creation and its multimodal distribution, including the production of parody, and content that pushes societal boundaries and self-reflection in ways that contribute to pluralism and inclusion. This has the potential to encourage the diversity of human expression and bring more people to participate in public debates on issues of public interest or ensure broader dissemination of content that might otherwise be limited to one form (text, for instance). Generative AI may aid the ability of users to create, re-use and distribute content, under the condition that the copyright and intellectual property rights are clearly established and respected, as well as the right to privacy, reputation and other rights that may be affected in this context.

46. **Personalised content:** Generative AI tools can enhance access to content and information of public interest by generating targeted and personalised messages, thus contributing to a better-informed public. Within public debate, Generative AI-powered chatbots or agents can provide voters with personalised informative content about current events, political developments and issues in text, voice or other formats. Such interaction may enhance political knowledge, improve access to informative content and facilitate public opinion formation, under the crucial condition that misuse is controlled.

47. **New tools for media, journalism, and fact-checking:** Generative AI can benefit democratic institutions of free speech, particularly the media, allowing them to develop new ways to inform and engage with the audience. Generative AI tools for aggregating, analysing, contextualising and summarising content can aid journalistic investigations, fact-checking, and media outreach.

***Structural implication 2*: Diversity and standardisation of expression**

48. **Loss of societal diversity and at scale homogenisation of expression**: Generative AI systems are based on statistical probabilistic systems. As such, they inherently produce outputs that align with the most represented training data in an unpredictable way or can mainstream certain ideas through advanced fine-tuning and guardrailing (see Figure 1, content moderation risks, Steps 4-5-6). While their impact may not be immediately noticeable on an individual level, their large-scale use can lead to significant societal consequences and implications for the diversity of human expression. One such consequence is the at scale homogenisation of expression, where unique or diverse voices risk being overshadowed by repetitive or statistically standardised content. This poses a growing challenge to individuals' freedom of expression, but for society at large where the distinct languages and cultures, or the expertise and reputation of those contributing to the diversity of the public debate (journalists, experts, individuals and communities), risk being standardised or diluted. The aggregate effect of such at scale homogenisation may threaten freedom of expression and pluralism[27].

49. **Standardisation of individual expression:** On an individual level, standardisation raises concerns about the diminishing diversity of expression in the private sphere, where personalisation risks narrowing perspectives rather than broadening them[28]. Empirical

[26] See examples of multimodal transfer between visual information and vocal information to help blind people in their everyday life, https://www.bemyeyes.com

[27] Effects on pluralism in augmented search span from content licensing deals to and political fine-tuning of conversational LLMs. See studies by Rutinowski et al. (2023) or Rozado David (2024).

[28] Hofmann et al. 2024 show that users can be discriminated against when using their own dialect when interacting with generative AI (through voice or writing), for example "Language models are more likely to suggest that speakers of African American English be assigned less-prestigious jobs, be convicted of crimes and be sentenced to death"

studies in real-world settings point to a loss of the diversity of human expression, by observing an at-scale standardisation of written or visual artistic expression. Concretely, participants asked to create content (e.g., product ideation tasks) with the assistance a Generative AI-based solution show a significant individual-level improvement of the ideas generated, while a substantial loss of lexical and content diversity of the formulations is actually registered (e.g., minus 41% of diversity)[29]. These kinds of empirical tests suggest how the at-scale effect of Generative AI use is yielding a standardisation of users' expression and of the ideas they convey, potentially leading to further long-term loss of cognitive capabilities to perform the same tasks. Similar effects of standardisation are observed beyond the domain of written or oral automated linguistic content creation in the visual domain[30].

50. **Lack of representativity of datasets:** Although Generative AI actors in the industry and in academia have been developing common practices in data collection, filtering, and pre-processing, the reality of Generative AI-based systems and their outputs often shed light on the fact that no training dataset is representative enough or covering all possible categories. There is thus a need for improvement and reflection on the impact that data collection criteria have on freedom of expression. Specifically, linguistic diversity, entailing also cultural diversity, as a precondition for broader representativity and inclusion, need to be tested by design[31] to ensure for example that low-resourced languages are not excluded and can also benefit from Generative AI in the context of freedom of expression.

***Structural implication 3**: Integrity of human expression and its attribution*

51. **Non-factuality, so-called Hallucination:** Predicting the most probable next words often conflict with facts and it is well documented that Generative AI-based systems routinely produce false answers or cite non-existent sources by statistically generating content to fill the gaps[32]. Although several technological refinements try to correct the inaccuracies of Generative AI augmented search, hallucinations pose a risk to an individual's right to access reliable information, one of the key elements of freedom of expression. The risk is also at societal level, where large scale use of Generative AI products can lead to widespread misinformation[33] and undermines trust and informational systems more broadly.

52. **Absence or blurring of information sources:** From the point of view of information accuracy, Generative AI-based tools are fundamentally different from search engines as they build content by statistically aggregating words to forge a new content consumption experience that has no identifiable sources or often inaccurate ones[34], thereby blurring the

[29] Dell'Acqua et al. 2023 Dell'Acqua, Fabrizio and McFowland III, Edward and Mollick, Ethan R. and Lifshitz-Assaf, Hila and Kellogg, Katherine and Rajendran, Saran and Krayer, Lisa and Candelon, François and Lakhani, Karim R., Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality (September 15, 2023). Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Paper, Available at SSRN: https://ssrn.com/abstract=4573321

[30] Automated image generation enabled by generative AI diffusion models (text-to-image) has an impact on human creativity in digital art. By examining 4 million artworks created by over 50,000 unique users of text-to-image generative AI tools, researchers observed the same dual effect : generative AI assistance in digital creation : While generative AI assistance to assist in digital creation enhances the appeal of the artworks by increasing the likelihood of receiving favourable peer evaluations per view by 50%, it also implies a significant decline in the average novelty of artwork content, alongside a reduction in the novelty of visual elements, as captured by pixel-level stylistic elements.

[31] See "SHADES: a Multilingual Assessment of Stereotypes in Large Language Models" study developing an LLM assessment tool (benchmark) on cultural stereotypes across 16 languages and 37 regions of the world by Mitchell et al. (2025), https://aclanthology.org/2025.naacl-long.600/.

[32] The challenge is that information generated by Generative AI is content that is structurally lacking the factuality of real information. More accurately said, it is a statistical representation of the linguistic distributions learned in training data. Generative AI based systems generate possible next words and sentences mimicking human productions, as such it can also be mis-information or dis-information.

[33] In line with the Recommendation CM/Rec(2022)12 on electoral communication and media coverage of election campaigns, and Recommendation CM/Rec(2022)11 on principles for media and communication governance, and the 2023 Guidance Note on countering the spread of online mis- and disinformation through fact-checking and platform design solutions in a human rights compliant manner, this Guidance Note considers both disinformation and misinformation. While both are understood as verifiably false, inaccurate or misleading content with potentially harmful effects for society, the difference is in it that misinformation spreads without a malicious intent, while disinformation is created and spread with an intention to deceive or secure economic or political gain. The spread of misinformation may be aided by technology and the way it is being used. Disinformation may as well spread faster and further due to the design or flaws in technology design but is a result of a strategic (ab)use of the technology and its affordances.

[34] A February 2025 study examined whether four leading AI assistants provide accurate responses to news-related questions and whether their answers faithfully reflected BBC News stories used as sources. Journalistic assessments

sources of information to an unprecedented degree. This context differs from the pre-AI information environment, which is based on discrete human artefacts such as articles or videos with associated authorship. This shift to Generative AI poses a risk to the right to access information and form opinions as it may diminish or remove people’s opportunity or ability to evaluate content based on sources.

53. **Dissociation from the author:** Generative AI may separate the work from the author, negatively affecting the author's right to impart information and potentially impacting trust in the informational ecosystem. Furthermore, it can dilute quality and even harm the reputation of the original author, for example, generating superficial summaries with wrong highlights. Authors have also warned about the risk of machines being prompted to appropriate their style, thus undermining and diluting the value and originality of their work and their voice[35].

54. **Advanced mimicking individuals’ personality:** Generative AI systems and their latest agentic development deepen the concern of a new era of deception and loss of attribution. AI agents’ systems can easily mimic an individuals’ personality with very little personal data input[36], and then replicate the values and preferences of the individuals to further act and accomplish digital tasks on the behalf of users. The easy access to resources that mimic the behaviours, attitudes, likeness and personalities of real people opens new possibilities for deception, loss of attribution and the dilution of freedom of expression. In addition, it raises the fundamental issue of whether individuals should have (a) the right to know if they are communicating with an AI or a human, or whether their messages are being received by a person or an AI, and (b) the right to know if they have been impersonated, and have redress mechanisms to require impersonations to be removed from training data sets and/or deleted from Generative AI products.

55. **Appropriation of likeness and deep fakes:** Generative AI tools misuse enables the appropriation of likeness, voice cloning, counterfeiting, impersonation, and the commoditisation of deep fakes. The creation and public dissemination of counterfeit or falsified content designed to impersonate an individual is often non-consensual and can evolve into a digital forgery. Deep fakes or other hyper realistic engineered audiovisual outputs enabled by Generative AI warrant high-risk to public discourse and information integrity overall, especially in the context of electoral processes. The potential for content manipulation, including spreading disinformation or impersonating candidates, journalists, and prominent public voices is a significant risk associated with Generative AI tools. Deep fakes are often used to distort public image and undermine the credibility of female voices in the public sphere.[37]

56. **Voice cloning:** In the sphere of voice cloning, the risk is higher for voices that are widely available online and in various repositories[38]. Cases of unfair and potentially unlawful cloning and selling of voices belonging to professionals in the voice industry have also occurred[39]. This raises concerns about the right of individuals - whose speech is accessible for Generative AI companies - to control its use and ensure its authenticity. Voice cloning incidents represent a large-scale dilution of individual personal expression amid fake and automatically generated statements[40]. Voice cloning, separate to other forms of multimodal impersonation of expression, represents additional risks to privacy, security and personal safety.

revealed that at least 20% of the responses contained significant inaccuracies, and up to 80% showed some form of accuracy issue. Additionally, 60% of the claims made in the AI-generated answers were, to some extent, unsupported by the sources they cited. https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf

[35] See: https://authorsguild.org/news/sign-our-open-letter-to-generative-ai-leaders/

[36] Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109.

[37] See in particular: GREVIO's General Recommendation No.1 on the digital dimension of violence against women and Protecting women and girls from violence in the digital age: the relevance of the Istanbul Convention and the Budapest Convention on Cybercrime in addressing online and technology-facilitated violence against women” (2021).

[38] See, for example, the case of Scarlett Johansson: https://www.npr.org/2024/05/20/1252495087/openai-pulls-ai-voice-that-was-compared-to-scarlett-johansson-in-the-movie-her

[39] https://www.bbc.com/news/articles/c3d9zv50955o

[40] One example is the high-profile case of the non-consensual appropriation of Scarlett Johansson’s voice by a Generative AI product and its implications for the value of the actress' voice and self-expression.

57. **Delegitimising or misusing prominent voices or outlets:** Generative AI may also be misused to hijack or undermine prominent voices, such as those of journalists, human rights defenders, or politicians, e.g. by generating and spreading false, inaccurate or misleading information about them or impersonating them. This can also affect media organisations (so-called "spoofing"). Blurring the lines between authentic and synthetic, accurate and fake content may worsen smear campaigns and harassment, particularly targeting female voices.[41] This may have a chilling effect on prominent, authoritative or critical voices, especially at risk given their potential reach and impact.

58. **Erosion of the information ecosystem and trust:** When produced and disseminated at scale, practices of false or mimicked online identities, used for deceptive purposes, create significant challenges for verifying and validating authentic communication. This further undermines information integrity and pluralism, as well as an individual's voice and self-expression, which can be diluted by deceptive artificial messages. The resulting confusion can weaken public trust and corrode the ecosystem of factual, reliable and diverse information. This risk derives from the current limitation of the technology but also from its intentional use in malicious ways.

***Structural implication 4**: Agency and opinion formation*

59. **Cognitive autonomy:** Generative AI systems and their use may also introduce new forms of disinformation, which function through continuous narratives, rather than isolated media artefacts, and are more easily scalable in production and distribution. As formulated in the Council of Europe Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes (Decl(13/02/2019)1), "sub-conscious and personalised levels of algorithmic persuasion[42] may have significant effects on the cognitive autonomy of individuals and their right to form opinions and take independent decisions", including of a political nature[43].

60. **Personalised persuasion:** Generative AI-based applications when used as search engines can also enable automated, personalised and interactive persuasion at an individual level. The fundamental difference from traditional search engines and the persuasive conversational mode of Generative AI, is that it can achieve persuasion and opinion shifts through simple text completion in a biased system[44]. Establishing an ongoing interaction akin to a relationship with a chatbot designed to achieve persuasive goals could lead to coordinated exposure to certain information over time. Examples of such persuasion leveraging users' conversation history, or hyper personalisation, have been documented in a range of use-cases from commercial marketing to political influence[45], as well as fully automated forms of online radicalisation, coercion, and emotional attachment, leading to some even taking their own lives[46].

61. **Loss of cognitive abilities:** Potential longer-term consequences derive from the frequent use of co-piloting tools that automate everyday cognitive tasks (e.g., co-writing, summarisation or other more complex tasks), leading to a loss of the cognitive ability to engage meaningfully with information and form opinions. Likewise, the extensive use of more autonomous AI agents that consume, process, and act on information on behalf of individuals can yield a loss of cognitive function and a weakening of critical thinking.

62. **Lack of AI literacy:** The engaging and enjoyable user experiences with Generative AI services, such as mainstream conversational agents or image generators, attract users who

[41] https://unesdoc.unesco.org/ark:/48223/pf0000387483
[42] See the experiments on latent persuasion in Jakesch et al., 2023.
[43] Bai, Hui, Voelkel, Jan, Eichstaedt, Johannes, *et al.* Artificial intelligence can persuade humans on political issues. 2023.
[44] Zeng, D., Legaspi, R. S., Sun, Y., Dong, X., Ikeda, K., Spirtes, P., & Zhang, K. (2024, April). Counterfactual reasoning using predicted latent personality dimensions for optimizing persuasion outcome. In *International Conference on Persuasive Technology* (pp. 287-300). Cham: Springer Nature Switzerland.
[45] Rogiers, A., Noels, S., Buyl, M., & De Bie, T. (2024). Persuasion with Large Language Models: a Survey. *arXiv preprint arXiv:2411.06837*.
[46] US Case Garcia v Character Technologies Inc (so-called Setzer Case where a 14 year old boy established a strong emotional attachment with a Character.ai designed upon a Games of thrones fictitious character) - https://socialmediavictims.org/wp-content/uploads/2024/10/FILED-COMPLAINT_Garcia-v-Character-Technologies-Inc.pdf

may not be fully aware of these models' underlying mechanisms, limitations, and risks, thus exposing individuals to the above cited risks without critical thinking, highlighting the need for increased literacy and education concerning Generative AI technology and its implications.

63. **Influence on individual opinion through latent persuasion:** Documented persuasiveness effects, opinion biases and users' overreliance on Generative AI output[47] arise from optimisation and design choices embedded in the later stages of Generative AI tools and products development (see Tool and Product Layer, Section 1). A subtle influence on end-users through tool design techniques, like prioritising user approval and satisfaction over accuracy or plurality (i.e. sycophancy), can be deceptive. These features leverage unconscious nudging techniques called "latent persuasion", leading users to adopt biases without realising it[48]. Large-scale experimental studies have documented how such techniques can induce opinion shifts on political topics or other forms of expression[49], thus eroding the autonomy and agency to form opinions and having profound implications at a society-level for the freedom to hold opinions.

64. **Large-scale automated opinion shift or manipulation:** Opinion manipulation through Generative AI can extend to critical areas such as disinformation and political speech, potentially having broader implications for democracy and the rule of law. These subtle but pervasive influences threaten informed decision-making and undermine foundational principles of the freedom to hold and form opinions through pluralistic debate.

<u>Structural implication 5</u>: Media and informational pluralism

65. **Efficiency gains in the media sector:** Generative AI-based applications may improve processes within media companies, such as marketing and distribution, automating tasks and generating story summaries tailored to various platforms and audience groups. It can also assist journalism through providing a set of tools to support research, documentation, analysis, enabling journalists to explore various angles of a story, as well as to verify, and create content[50]. This could help alleviate economic pressure on media companies and repetitive tasks for journalists and create a positive effect on the media landscape.

66. **Impact of biased datasets on pluralism:** If Generative AI models are trained on partial or biased data sets, their outputs may amplify pre-existing biases, and undermine pluralism - that is, the diversity of opinions, perspectives, and sources that reflect the plurality of society. This includes linguistic diversity and raises concerns about preserving especially underrepresented languages in the digital and AI-mediated future, including the potential of individuals to express themselves or receive information using Generative AI applications. Such risks have led some countries to offer their language and data sets, simply to ensure representation in Generative AI models and developments.[51] Empirical evidence already shows various dimensions of amplifying stereotypes and gender biases[52]. There is also a risk of mainstreaming majority voices, making minority voices even less visible[53], as well as

[47] Steyvers, M., Tejeda, H., Kumar, A. et al. What large language models know and what people think they know. Nat Mach Intell (2025). https://doi.org/10.1038/s42256-024-00976-7
Researchers at UC Irvine conducted a study on three publicly available LLMs (GPT-3.5, PaLM2, and GPT-40) and found that users consistently overestimate the accuracy of LLM outputs and tend to rely on longer explanations more (i.e. "length bias") : The inability of users to discern the reliability of LLM responses not only undermines the utility of these models but also poses risks in situations where user understanding of model accuracy is critical.

[48] Jackesh et al., 2023 and Zeng, D., Legaspi, R. S., Sun, Y., Dong, X., Ikeda, K., Spirtes, P., & Zhang, K. (2024, April). Counterfactual reasoning using predicted latent personality dimensions for optimizing persuasion outcome. In *International Conference on Persuasive Technology* (pp. 287-300). Cham: Springer Nature Switzerland.

[49] Rogiers, A., Noels, S., Buyl, M., & De Bie, T. (2024). Persuasion with Large Language Models: a Survey. *arXiv preprint arXiv:2411.06837*.

[50] https://charliebeckett.medium.com/what-we-have-learnt-about-generative-ai-and-journalism-and-how-to-use-it-7c8a9f5e86fd

[51] See, for example: https://openai.com/index/government-of-iceland/

[52] Different language models were shown to be significantly more likely to generate cover letters with less formal tone (e.g., sentence structure and phrasing) for women compared to men Additionally, the lexical choices often reflect stereotypes and gender bias). See Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). " Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. arXiv preprint arXiv:2310.09219.

[53] Campbell, C. 2024. Automated Journalism at the Intersection of Politics and Black Culture: The Battle against Digital Hegemony. Lanham, Maryland: Rowman and Littlefield

mainstreaming radical and vocal minority voices, which could be overrepresented in training data. It is also possible that, through data strategies, the opinions and ideas that the owners of AI tools support ideologically, will be amplified, at the expense of others.

67. **New gatekeepers and economic disruption in the information ecosystem:** The rapid and widespread adoption of Generative AI-based augmented search applications as information sources is establishing new intermediaries between the media and their audiences and may disrupt the reach and economic viability of the media. If Generative AI-based systems could rely on high-quality, up-to-date content to produce accurate outputs, this could create a potential revenue stream for the media and journalism sector. However, current practices often involve the unauthorised use of media content - which is expensive to produce - for model training and output generation. This raises concerns about the economic sustainability of the media and other creative industries. Even in cases where remuneration or licensing deals are established, they lack transparency, and the preference for major publishers from bigger markets over smaller ones further raises concerns about the lack of linguistic and cultural representation, as well as access to diverse and local information. The concern is ultimately about safeguarding media pluralism[54], as a corollary of freedom of expression and the integrity of the information sphere.

68. **"Audience of one":** Generative AI is further shifting the diffusion of information towards a one-to-one paradigm. An "audience of one" stands for an information environment where everyone interacts with Generative AI-powered information separately, and receives hyper personalised and unique content, which will not be received by anyone else. This has the potential to create a "bubble of one", where individuals are fed by personalised streams of information that reinforce existing personal beliefs and biases, even misperceptions. This way, the very core notion of a shared informational space is diluted, with numerous consequences for democracy and freedom of expression and specifically the right to hold opinions. Thus, making individuals more vulnerable to manipulation and less likely to agree on basic facts. In the long term this can reinforce the ongoing process of societal fragmentation of the informational space and polarisation.

Structural implication 6: Market Dynamics

69. **Potential market dynamics:** The market dynamics of the Generative AI technology lifecycle are fast evolving. While sharing some characteristics, they are in many ways different from the dynamics and network effects of online platforms. They are shaped by some key factors like access to data, talent, capital and computing power; each factor being subject to its own market dynamics, with the presence of only a small number of actors at some layers of the Tech Stack. This presents not only competition challenges[55] but can lead to significant concentration with undue implications for freedom of expression at each layer of the Generative AI Tech Stack.

70. **Lack of inclusive and accountable AI design:** The design, development, optimisation and deployment of Generative AI can reflect the political and economic interests of single actors in the Generative AI Tech Stack or be driven by a specific business model, rather than prioritising societal benefits or acting in the public interest. Furthermore, when Generative AI optimisation and content moderation, as a tool and product layer design choice, lacks inclusivity, meaningful participation of relevant rights-holders, oversight and accountability, there is a risk of undue influence over freedom of expression.

[54] Understood in a broad sense along the four dimensions as operationalised by the Media Pluralism Monitor: (i) Fundamental Protection (of fundamental rights to freedom of expression and access to information, status and safety of journalists), (ii) Market Plurality (considering both digital and traditional markets, content production, distribution, and consumption), (iii) Political Independence (of a newsroom, but also of a wider media and information structure and resources), Social Inclusiveness (access and representation of various societal groups, especially those in vulnerable conditions) https://cmpf.eui.eu/media-pluralism-monitor/

[55] UK Competition and Markets Authority technical report on competition implication of AI Foundation Models (dated 16 April 2024) https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf; French Competition and Market Authority's report in 2023 : https://www.autoritedelaconcurrence.fr/fr/communiques-de-presse/intelligence-artificielle-generative-lautorite-rend-son-avis-sur-le. The EU and the US have ongoing investigations.

71. **Concentration at the Foundation Layer:** The current layering of the Generative AI Tech Stack reinforces the concentration of market power at the foundation layer. This initial layer is characterised in the current technological advancement of Generative AI by a high concentration of the three key success factors: talent, data, and computational investments. Currently, this configuration strengthens the market power of incumbents in the field and creates structural dependency for actors at the other layers. A natural mitigation of the observed concentration is linked to the emerging trend of developing smaller specialised models or to build multi-models composite systems to better achieve complex tasks through AI Agents, and open-source models rapid developments. Notably, open source could offer varying levels of more transparent and valid alternatives, but also comes with its own risks, e.g. when open-source models are not appropriately vetted or maintained.

72. **Tool Layer and specific design risks for freedom of expression:** Market concentration is less evident at the Tool layer as an increasing number of smaller entities are working to adapt foundation models to specific tasks. Infrastructure investment and technological expertise are lower than the one needed to innovate and be competitive at the Foundation layer. Major investments at this stage are in data quality (and not quantity) to perform model instruction and adaptation (see Steps 4 and 5, Figure 1). However, while there is more diversity of actors at this layer, actors can be seen in a position of structural dependency from the foundation layer. Current trends in Generative AI technological development move towards the use of small LLMs and the open-source developments may alleviate concentration and concerns about transparency.

73. **Specific design risks at the Tool layer:** The content moderation policies implemented at this layer have profound implications for freedom of expression, potentially undermining the rule of law, and call for a specific oversight. As presented in section 1 and in this section, the exercise of fine-tuning guardrails and filters, and other measures that direct tool performance, like content alignment with human preferences, can cause unjustified interference with the right to freedom of expression. In cases of vertical concentration across the different layers of the Tech Stack, dominant actors can exert significant control over how expression is standardised and controlled or how content moderation is performed. This may lead to content moderation practices that are disproportionate to their intended benefits and risk undermining the rule of law.

74. **Product Layer and user dependence:** Vertical concentration of market actors across the Generative AI layers of Tech Stack and their end-user data capture to design hyper-personalised products contribute to the lack of viable alternatives particularly at the product layer. Hence, the design of the Generative AI applications influences, nudges and drives the behaviour of its users to be dependent on the product and/or become (over) reliant on product outputs. The currently impossible portability of user interaction history to enable moving from one generative AI powered product to another in a frictionless manner is representing a further limiting effect on freedom of expression. The lack of transparency of design and implementation at the product layer makes it harder to observe and mitigate potential freedom of expression risks or for regulators to hold those actors accountable.

SECTION 4 – GUIDELINES

75. Member states have a positive obligation to foster an environment where freedom of expression can thrive. Securing the right to freedom of expression when mediated through Generative AI technologies and applications is vital to ensuring an enabling environment which promotes and protects freedom of expression in all its dimensions. As highlighted in Section 1, benefits and risks for freedom of expression are present across the current Generative AI Technological Stack. Effectively reaping the benefits and mitigating the risks requires a clear understanding of what is at stake for freedom of expression (Section 3). Considering the six structural implications across the layers and actors of the Generative AI Tech Stack as a guiding framework is essential to create a favourable multi-stakeholder dialogue which promotes and protects freedom of expression.

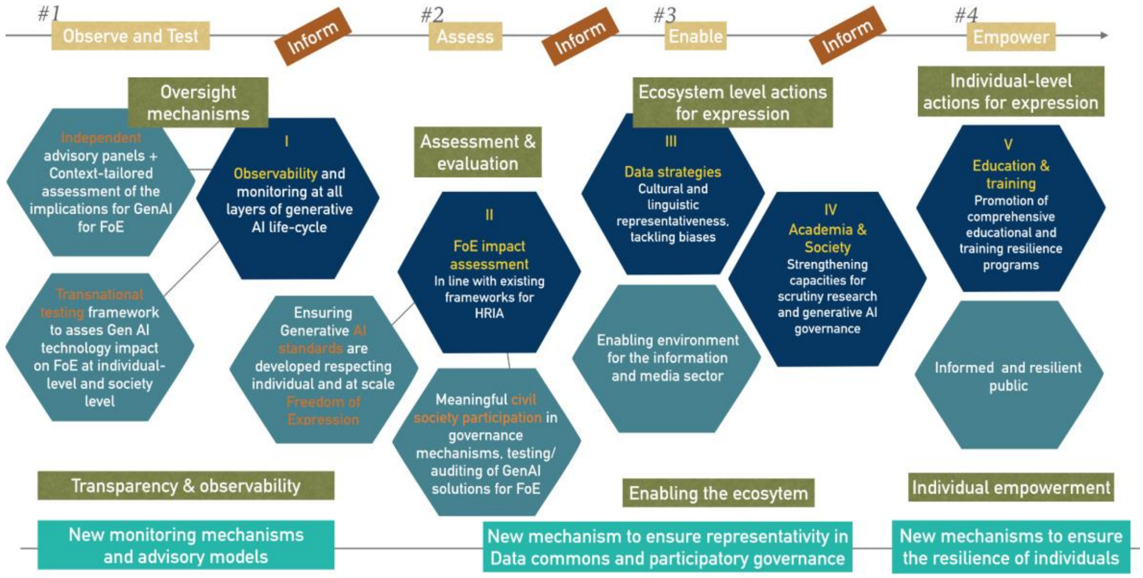

#1
Observe and Test
Inform
#2
Assess
Inform
#3
Enable
Inform
#4
Empower
Oversight mechanisms
Independent advisory panels + Context-tailored assessment of the implications for GenAI for FoE
I
Observability and monitoring at all layers of generative AI life-cycle
Transnational testing framework to asses Gen AI technology impact on FoE at individual-level and society level
Ensuring Generative AI standards are developed respecting individual and at scale Freedom of Expression
Assessment & evaluation
II
FoE impact assessment
In line with existing frameworks for HRIA
Meaningful civil society participation in governance mechanisms, testing/ auditing of GenAI solutions for FoE
Ecosystem level actions for expression
III
Data strategies
Cultural and linguistic representativeness, tackling biases
IV
Academia & Society
Strengthening capacities for scrutiny research and generative AI governance
Enabling environment for the information and media sector
Individual-level actions for expression
V
Education & training
Promotion of comprehensive educational and training resilience programs
Informed and resilient public
Transparency & observability
Enabling the ecosytem
Individual empowerment
New monitoring mechanisms and advisory models
New mechanism to ensure representativity in Data commons and participatory governance
New mechanisms to ensure the resilience of individuals

Figure 2: Detailed actionable steps of the Guidance Note on the implications of Generative AI for Freedom Expression.

76. Member states should take proactive steps to ensure that Generative AI applications, their design and use upholds freedom of expression and mitigate risks to it. The following recommendations aim to provide member states with guidance on how this can be achieved. They are divided into four action areas:
    A. **Observe** the impact of Generative AI applications and technology on freedom of expression through **robust oversight and testing mechanisms** evaluating its potential positive and negative effects. This approach will enable transparency measures, help identify biases, and foster responsible data governance.
    B. **Assess** Generative AI systems through **ongoing risk assessments** including systematic, tailored and inclusive freedom of expression impact assessments and due diligence in public procurement.
    C. **Enable** the full exercise and protection of **freedom of expression rights**, including by strengthening socio-technical standards.
    D. **Empower** relevant stakeholders by adopting a wide range of measures aimed at **awareness-raising and participatory approaches** to risk governance (including citizens' assemblies), education, research, publication of impact assessment findings, facilitating user choice and other international cooperative approaches**.**

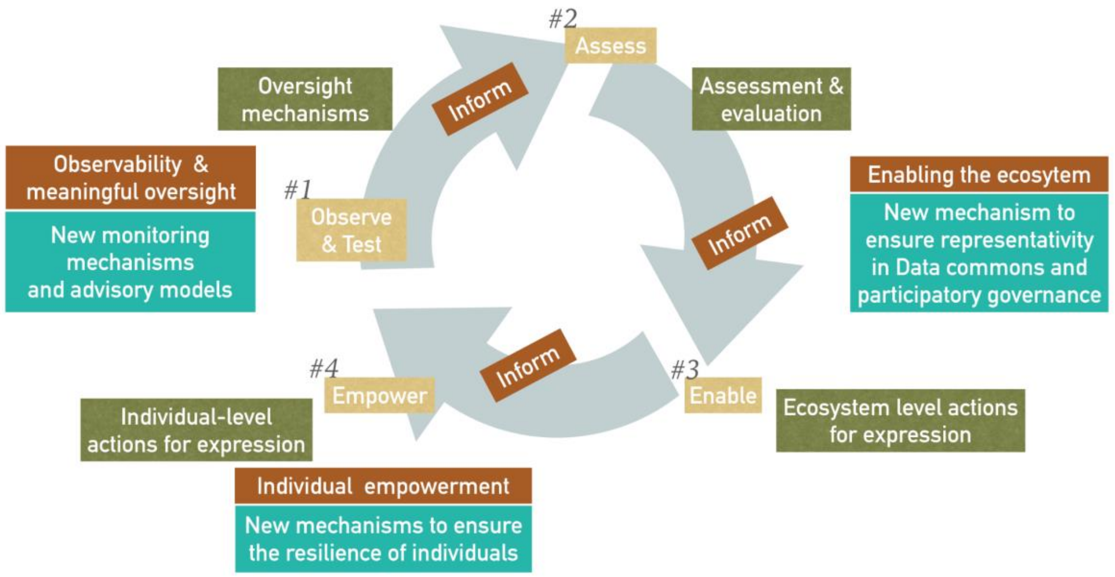

#2
Assess
Oversight mechanisms
Inform
Assessment & evaluation
Observability & meaningful oversight
New monitoring mechanisms and advisory models
#1
Observe & Test
Enabling the ecosytem
New mechanism to ensure representativity in Data commons and participatory governance
Inform
#4
Empower
Inform
#3
Enable
Ecosystem level actions for expression
Individual-level actions for expression
Individual empowerment
New mechanisms to ensure the resilience of individuals

Figure 3: The Observe, Assess, Enable and Empower feedback-loop for Policy action on the implications of Generative AI for Freedom Expression.

77. The above action areas (presented in more detail below) are intended to present policy makers with fundamental building blocks to safeguard freedom of expression. As each of the action areas are progressed, corresponding follow up actions are required to inform and provide feedback. Feedback should detail the particular implications for freedom of expression (and also impacts to democracy, the rule of law and other human rights) that have been observed and assessed can be made publicly available and reported in a way which is accessible to a wide variety of actors. By taking informed actions, relevant stakeholders can enable a favourable ecosystem for freedom of expression to flourish and to empower individuals to become more resilient.

OBSERVE

78. Observing and monitoring the positive and negative effects of Generative AI systems on the exercise of freedom of expression by individuals and groups is the most fundamental precondition to understanding how Member states can promote freedom of expression and ensure its proper exercise or any mitigation action. Being able to observe and monitor the complex and rapidly evolving implications of Generative AI for freedom of expression requires to focus on three fundamental dimensions to achieve meaningful **understanding, oversight** and **transparency**: (1) the ever-evolving technology, (2) its rapidly adopted applications, and (3) the underlying market dynamics.

79. To identify and monitor benefits and structural risks in real world use cases, Member states should establish competent advisory panels and **meaningful monitoring mechanisms at a national and international level,** based on procedures and practices that gather information and greater insights into the workings of Generative AI Tech Stack and its actors across borders.

80. Member states should foster **effective international cooperation and coordination** of observatories to ensure **observations concerning the impacts on freedom of expression associated with Generative AI are shared** (including that from market dynamics). **International collaboration** will be key to addressing internationally recognised observations. To ensure cross-border and multi-stakeholder engagement, Member states should consider advisory models involving multilateral authorities, private sector, independent experts, affected users, civil society organisations, academia.

81. To identify and to effectively respond to Generative AI challenges to Freedom of expression, Member states should design and set up meaningful **observation mechanisms** that

systematically test, monitor and provide an oversight mechanism for the impacts on freedom of expression with careful consideration for the following:

a. having access to the **relevant expertise** with the necessary technological background and human rights knowledge and independence;
b. acting in the **public interest** and with **legitimacy**;
c. ensuring **inclusion of relevant expertise** from a wide range of stakeholder perspectives, making the participation possible of the private sector, impacted users, civil society organisations, academia and intergovernmental organisations;
d. providing public access to **findings to provide important freely available information and to foster an ecosystem transparency**;
e. having permanent testing environments, fully resourced with competent professionals and tools to assure **continuous monitoring**;
f. foster effective cooperation and coordination between relevant national and international regulators and appropriate bodies;
g. ensuring that the observatories' structure, support, operations, and funding uphold their **independence** and **maintain public trust**.

82. Member states should facilitate the publication of detailed findings of testing conducted through the observatories of identified risks and mitigation strategies for freedom of expression. By making the findings readily available and accessible through observation and monitoring reports, increases human oversight of Generative AI systems and transparency and raises awareness amongst stakeholders and end-users, acting **as a means of epistemic counterpower.**

83. Member states should consider the **professionalisation of Generative AI testing** and requiring testers to have the necessary technical expertise and human rights knowledge to ensure testing and observation of freedom of expression impacts is consistent and of high-quality.

ASSESS

84. Member states should advocate for the inclusion of the implications on freedom of expression within **human rights risk and impact assessment for Generative AI systems and applications**. Existing mechanisms, such as the Council of Europe Methodology for the Risk and Impact Assessment of Artificial Intelligence Systems from the Point of View of Human Rights, Democracy and the Rule of Law (HUDERIA Methodology),[56] provide a solid basis to further develop a targeted, inclusive, and consistent approach specific to the Generative AI implications for freedom of expression.

85. Human rights risk and impact assessments must be **systematic, iterative, robust, and flexible** in covering the entire Generative AI Technology Stack, end to end, to effectively assess the risks that Generative AI poses to freedom of expression. The following key considerations should guide this approach:

a. **Risk and impact assessment and resultant mitigation measures** should be co-developed by member states and actors operating within the Generative AI Tech Stack as well as those directly impacted and affected by them. To this end, member States should consider establishing protocols for **participatory conduct** of **freedom of expression due diligence** in all new Generative AI public procurements, to provide an opportunity for meaningful engagement of civil society and the public in assessing societal and individual impacts on freedom of expression.
b. **Co-development of documented and auditable trail for impact assessment with actors operating in the Generative AI Tech Stack, including for example details** on intended purpose, justification for safeguards, applied optimisation and fine-tuning, data and model choices, meaningful stakeholder engagement and mitigation strategies.
c. **Accessible and meaningful information and explanations** about how Generative AI systems operate, their implications for freedom of expression, and the safeguards in place

[56] The HUDERIA Methodology was adopted by the Council of Europe's Committee on Artificial Intelligence (CAI) at its 12th plenary meeting, held in Strasbourg on 26-28 November. It will be complemented in 2025 by the HUDERIA Model, which will provide supporting materials and resources, including flexible tools and scalable recommendations.

to mitigate risks should be made publicly available and accessible to citizens and civil society.

86. **Specialised training** should be required for those responsible for conducting freedom of expression risk and impact assessments whether they be from the public or the private sector. Relevant standards and case-law of the European Court of Human Rights should inform these trainings. Expertise can be drawn from the Council of Europe, Human rights organisations, and Equalities Bodies who have historically through other cooperation activities been able to provide an exchange of professional views, opinions and experience that could play a key part in upskilling specialised assessors. Member states should promote access to appropriate human rights and legal training for designers and developers of Generative AI tools which set parameters on how end products and applications perform, especially when being used in judicial systems and public services and infrastructure.

87. In assessment and training, particular attention should be given to the impact of Generative AI on **individuals and groups** in a position of **vulnerability** such as children, persons from marginalised communities, persons with disabilities and those in precarious physical, emotional, financial or psychological situations. Vulnerable persons or groups may be more susceptible to mental health impacts, opinion shifts, latent persuasion or social inequalities. Women may be more susceptible to AI-driven harassment, or technology-facilitated exploitation, to the publication of personal and often sensitive information on the internet usually with malicious intent (known as “doxing”), and gender-based violence through Generative AI impersonation and deep fakes.[57]

ENABLE

88. Any strategy to maximise the benefits of Generative AI and reduce its risks to freedom of expression depends on an enabling environment - one where member states actively support the development of a Generative AI ecosystem that promotes human rights. Creating an enabling environment requires Member states to:
    a. **Work towards building a coordinated international oversight and observatories networks.** This network should include diverse disciplines and sectors of society and support the need to observe and assess Generative AI’s impact on freedom of expression across borders.
    b. **Strengthen the capacity of academia and civil society** by providing structured support for the important **independent research**, **capacity building and awareness raising.**
    c. **Protect credible and reliable information sources** and enable the continued ability to obtain authentic information from multiple sources.
    d. **Incentivise investment in the development and adoption of socio-technical standards**[58] to ensure that Generative AI is (1) developed to promote and to protect freedom of expression by design and by proactively seeking to mitigate against systemic and structural risks, and (2) interoperable.

89. **Protection of authentic information sources** entails Member states providing conditions for **an independent and pluralistic media ecosystem and allowing journalism to play an essential public watchdog role**, as well as fostering new forms of public interest content production, access and distribution. Given the potential negative implications of Generative AI and the broader digital transformation to the visibility and economic viability of journalism, Member states should consider supporting the development **of public service digital information infrastructures** as an alternative to commercially driven infrastructures and applications.

90. Member states should enable **interoperability through rights-respecting industrial standards** that enhance transparency and observability, which further equip independent

[57] See in particular: GREVIO's General Recommendation No.1 on the digital dimension of violence against women and Protecting women and girls from violence in the digital age: the relevance of the Istanbul Convention and the Budapest Convention on Cybercrime in addressing online and technology-facilitated violence against women” (2021).

[58] Leveraging international standards, such as ISO, IEEE, CEN/CENELEC could help co-develop essential socio-technical standards for testing and benchmarking of Generative AI tools and applications for freedom of expression impacts.

assessment and testing in line with freedom of expression, support oversight and contribute to a more open, innovative and competitive digital ecosystem grounded in human rights.

91. Member states, in collaboration with the private sector and civil society, should consider **investing in data strategies** fostering the **development of accessible, diverse and representative public data sources** that **support freedom of expression and** information pluralism and responsible Generative AI governance across the Tech Stack. This could include the creation of dedicated data spaces for certain areas of application to resolve the data-related concerns presented in Section 3. Such **data sources allow for training, evaluating, validating, and verifying Generative AI outputs**. Member states should pay particular attention to data sources that are relevant in relation to freedom of expression, media diversity and information pluralism, to safeguard democracy, rule of law and equality. Member states should maintain access to diverse and inclusive data spaces and data for training Generative AI pursuing the following goals: (1) limit the risk of standardisation of expression and of undermining the rule of law, (2) minimise unwanted bias and discrimination, and (3) to take actionable measures that ensure a certain degree of national technological sovereignty.

92. By enabling **greater transparency on data collection, usage, and access**, such public data sources can enhance transparency in generative AI development, design and optimisation. Making these data sources accessible for scrutiny and audits by independent entities - including regulators, civil society organisations, academia, and technical experts - can improve responsible development. This approach fosters a responsible use of data in Generative AI, balancing innovation with data protection, and end-user privacy amongst other human rights considerations, and mitigating against the distortive effects of generative AI on opinions, the potential for standardisation of the end-user human expression and for polarisation by AI assisted and mediated outputs.

93. Member states, together with actors operating within the Generative AI Tech Stack, should take steps to promote freedom of expression by improving the **identification and proposals for mitigation of biases and disparities in data, especially in training and fine-tuning foundation models** so they are more inclusive. Tackling disparities in data representation or increasing transparency on data sources used at the Foundation and Tool layers, and fostering information pluralism, will help to **address linguistic and cultural diversity gaps** which have the potential to have exclusionary effects for those whose language is little represented**.**

94. Member states should consider **incentivising measures to encourage the availability of more diverse Generative AI-powered product choice and viable technical alternatives.** Such measures could include requiring portability of users' interaction data and minimum interoperability requirements. This could counter unhealthy concentration dynamics in the market, end-users' data capture, hyper personalisation side-effects and thus encourage users to exercise their free choice amongst diverse GenAI applications.

EMPOWER

95. For empowerment to be effective member states should enable a multi-stakeholder approach to:
    a. strengthen **education and literacy in Generative AI** and freedom of expression among other human rights,
    b. improving **avenues for seeking and obtaining redress** where Generative AI harms to freedom of expression have occurred,
    c. develop **regulatory and non-regulatory approaches** to help companies and users to incentivise responsible ecosystem behaviours,
    d. Participate in an **open dialogue** among stakeholders across various intergovernmental fora, such as the Council of Europe. This dialogue should involve industry, academia, civil society, and public administrations at local, regional, national and international levels.

96. Member states should draw on lessons from the media literacy landscape to create **accessible public resources on Generative AI**, aimed at improving understanding of its

implications for freedom of expression. These efforts should raise awareness across diverse demographics, social groups, and within the public sector.

97. Member states should promote **comprehensive education both in school and at the workplace** by giving access to cross-functional training for lifelong learning concerning both the workings of Generative AI at different levels of the technological stack, and their risks and impacts on freedom of expression. Such training is especially crucial in judicial and public service sectors that are integrating or operating with Generative AI tools and products.

98. Member states should **make it evident and more accessible to individuals and groups how to obtain redress** when their freedom of expression is unduly restricted by the Generative AI design or use, where the scope of consumer protection ends. To this end, member states should work with and support **human rights and civil society organisations and academia** to provide means of obtaining evidence to demonstrate how Generative AI implications for freedom of expression occur, and to disseminate information about available redress mechanisms to those who could be negatively impacted. To this end, member states should consider establishing sustainable funding mechanisms for organisations operating in this field.

99. **Member states should promote a package of redress mechanisms**, both for individual users and business users, as well as collective redress for societal level harms. Redress could be obtained across the whole Generative AI technological stack - if observability and transparency measures are not sufficient to distinguish one actor from another - or at each layer of the Generative AI Tech Stack. Redress mechanisms can include a means to:

a. a user ceasing use of a generative AI product,
b. a regulator suspending a generative AI product from the market until appropriate corrective actions are implemented,
c. find an alternative and make an informed choice, including the possibility of a publicly funded option designed to serve broader public interests,
d. access and download one's information from the generative AI product,
e. ensure that individuals can obtain a meaningful explanation of how generative AI was used and access evidence of how the system operates,
f. get access to resources to enable users to overcome barriers to seeking legal and human rights help from appropriate ombudspersons, public authorities, human rights bodies, tribunals or courts, especially where the potential for freedom of expression harms can be disempowering.

100. Member states, in collaboration with civil society, should support actors across the entire Generative AI Tech Stack in enhancing transparency, expanding user choice, incentivising responsible market behaviour, and fostering international coordination to share insights on impacts to freedom of expression. A range of regulatory and non-regulatory tools may be employed to address harmful ecosystem dynamics. These could include sector-specific codes of practice, regulatory warnings, and the publication of risk and impact assessments, as well as performance metrics relevant to individuals seeking redress for freedom of expression violations.