AI as a Public Good: Ensuring Democratic Control of AI in the Information Space

POLICY FRAMEWORK

FEBRUARY 2024

Forum on
Information
& Democracy

CONTENTS

A ROADMAP FOR AI IN THE PUBLIC INTEREST
BY LAURA SCHERTEL MENDES & JONATHAN STRAY

Co-chairs of the Working Group

Six months ago, when we were asked to co-chair the Working Group on Artificial Intelligence and its Implications for the Information and Communication Space, set up by the Forum on Information and Democracy, and embarked on this ambitious process to make a substantive and democratic contribution to the global debate on AI and its regulation, many questions remained unanswered.

I, Jonathan, a computer scientist and a journalist, have been studying AI systems' influence on outcomes including informed citizens, user well-being, and most importantly polarization and conflict. With my colleagues at the Center for Human-compatible AI at UC Berkeley, we are investigating what problems arise and what solutions exist. A huge fraction of the information that all people see is now selected by increasingly sophisticated algorithms. What biases do these ubiquitous systems have, and how will what they show us escalate or de-escalate our disagreements? Furthermore, how can bad actors manipulate these systems to deceive or divide us, through synthetic media or networked persuasion? We need to do better than just letting things happen, through a combination of careful design principles and regulation that guarantees responsiveness to public needs. And we need policy mechanisms to let external researchers study the internal workings of commercial systems, so we can all understand what these hugely influential machines are actually doing to individuals and societies.

I, Laura, a Professor of Law, have been deeply involved in translating ethical principles into AI regulation. As the rapporteur of the Brazilian Jurists Commission I advised the Brazilian Senate on the approach to take in regulating AI, elaborating Brazil's first draft AI Law. We aim to put in place comprehensive AI governance to ensure that AI technologies are developed and used responsibly and ethically while minimizing potential harms and risks. We want to create, through regulation, genuine accountability mechanisms external to companies and accessible to society. We recognise that this can take different forms in different countries, to cater to diverse cultural backgrounds, as we discussed in our global Working Group.

This report is the result of six months of intensive and inspiring discussions, testing ideas with experts from various disciplines and corners of the world. Our Working Group meetings included dedicated sessions on a wide range of topics including accountability regimes, technical mitigations, authenticity standards and policy options. These discussions have led us to a roadmap for ensuring that AI is developed, deployed and used in the public interest. By framing AI as a public good, we are advocating for a shift in priorities. AI systems must be safe, fair, and reliable if they are to be an innovation that benefits all people and promotes sustainable development.

The democratic governance of the information and communication space is a precondition for democracy itself. The responsible use of AI can promote this democratic governance or, on the contrary, mitigate it. The series of measures proposed in this policy framework enable democratic control of AI in the information space.

Our recommendations are addressed to actors including AI developers, deployers, and governments, calling upon them to implement concrete and practicable measures to build more ethical, inclusive and responsible AI systems. The people and organizations powering the AI revolution can be at the forefront of technological developments that serve the interests of citizens.

The recommendations are formulated in a flexible way so that they can be implemented by different actors in very different contexts and countries. At the same time, in response to the demand for guidelines that go beyond general principles to guide practice, they are concrete and propose specific measures.

We put citizens at the heart of our concerns, calling for proactive consultation, accountability systems and redress mechanisms. We are combining a risk-based approach to AI regulation with a rights-based approach that safeguards the right to be informed, to not be discriminated against, to receive an explanation and to challenge a machine-generated outcome.

Those of use who are researchers and civil society actors play a particularly important role in studying these systems, scrutinizing them to verify compliance but also to inform policy-making and the development of more trustworthy systems. Yet, AI developers are reluctant to provide access to their systems and training data. Only through appropriate regulation can we empower external stakeholders to fulfill their contributions to a technological development that serves democracy.

We would like to thank our colleagues in the Working Group for their valuable insights and contributions, all of the many experts who donated their time and knowledge, as well as the extraordinary efforts of the team leading and drafting this policy framework: Katharina Zügel, Viviana Padelli, Lia Chkhetiani and Kaye Celine Palisoc.

If AI development and use continues as it currently does, it poses major challenges to the information environment that powers democratic processes. As an American and Brazilian, we look with worry toward the upcoming elections in our countries. We are on the verge of a major shift in the AI governance landscape from ideas to regulation. It is time for States to act, and our roadmap is intended to guide policy-makers in defending democracy.

DEFENDING DEMOCRACY IN THE FRONTIER OF ARTIFICIAL INTELLIGENCE
BY MICHAEL BĄK

Executive Director, Forum on Information and Democracy

The trajectory of technological evolution – including that of artificial intelligence – is not inevitable, despite the oracle-like pronouncements of tech billionaires and executives. Nor must it be built to conform to the priorities of private interests, capital, and shareholders. Artificial Intelligence is a public good. It must serve us not only as consumers, but first and foremost as citizens.

Our democratic institutions have a responsibility to shape and guide the evolution of AI in a responsible direction, one that conforms to the shared values of our democracies, respects the agency of people everywhere, and strengthens our fundamental human rights, including that of the right to reliable information.

Launched in 2019 and currently with 52 state signatories, the International Partnership for Information and Democracy works to ensure that our democratic institutions govern the global information and communication space by democratic rules and norms, developed in partnership with civil society. Our organization, the Forum on Information and Democracy, is the civil society led entity that brings this vision to life. And we are focused on preventing, limiting, and mitigating stress placed on our democracies due to unrestrained tech, including artificial intelligence.

Artificial intelligence presents an unprecedented transformation in how we create, disseminate and consume information. It decides what you see and what I see; and that we don't see the same things. These systems enable anyone to easily create and disseminate information, yet they are often biased, discriminate against specific groups, or hallucinate. AI systems can also easily be abused by malicious actors that seek to deceive citizens, influence political processes and sow doubt on the facts that form the bedrock of democratic discourse.

Our recommendations aim to pre-empt and prevent these harms and steer technological innovation as a public good, in a direction that is in the public interest.

We must not commit the same errors as in the past, where social media and tech companies decided the rules of the game, set the agendas, determined which harms mattered (and where), and captured the policy narratives. This resulted in too much corporate apology and, sadly, substantial harm to our communities, our democratic institutions and to our agency as citizens.

We believe that artificial intelligence can take a different, more enlightened, meaningful path. AI technology can move forward on a path guided by democratic oversight, constantly assessed and improved through civic leadership and inclusive participation.

We must ensure inclusive frameworks and mechanisms that allow citizens to ensure AI systems are developed and deployed in the interests of our diverse world. And this can only happen with transparency, accountability, and democratic oversight.

This report presents recommendations to achieve this. And this can be done while encouraging thriving innovation.

On 28 September 2023, on the occasion of the International Day for Universal Access to Information, we gathered fourteen leading experts from various regions and disciplines to start an inclusive and consultative process to elaborate critically important policy recommendations to governments and AI companies. Co-chaired by Laura Schertel Mendes, Lawyer and Professor of Civil Law at IDP in Brazil, and Jonathan Stray, Senior Scientist at the Berkeley Center for Human-Compatible AI in the United States the group represents a brains trust of significance. Nearly half the group comes from the Global Majority, half are women, and they all represent a wealth of diversity of lived experience, academic rigor, and commitment to responsible and ethical AI.

The rapporteurs team led by Viviana Padelli, Economist and Public Policy Expert, and Katharina Zügel, Policy Manager of the Forum, interviewed and consulted more than 150 people from diverse backgrounds around the world in this process. With the guidance of the working working group, they have articulated urgent actions for the world's policymakers, AI companies, and civil society as we all march forward into our AI future.

This report covers a rich arc of considerations and active measures that policymakers and AI companies can take to ensure AI serves our democracies and our citizens.

We deeply appreciate the contributions of the entire Policy Working Group and our partners around the world who have contributed significant effort and dedication to ensuring we get AI right – for the sake of our democracies, our citizenship and our future generations.

Time is of the essence. Perhaps no technology in the history of humanity has developed at such exponential speed. And we must keep pace.

KEY RECOMMENDATIONS

AI companies and entities can proactively build more inclusive and trustworthy AI systems by:

- Setting up an **inclusive and participatory process** that at a minimum includes equitable, sustained and substantive participation by independent researchers, civil society, and affected communities, to determine the **rules and criteria guiding dataset provenance and curation, human labeling for AI training, alignment, and red-teaming**. These rules and criteria should be publicly available.
- Implementing **risk mitigation and output moderation measures**, including **notice and action mechanisms** for capturing user feedback, **complaint mechanisms**, collaboration with **trusted flaggers and fact-checkers,** and continuous **red-teaming** to tackle identified risks.
- Providing users with an easy and user-friendly opportunity to choose **alternative recommender systems** that do not optimize for engagement but build on ranking in support of positive individual and societal outcomes, such as reliable information, bridging content or diversity of information.
- Implementing a policy where both content and users must acquire a "**right of recommendability**" before getting promoted or seen in feeds. This right should be granted based on a valid cryptographic signature linked to trusted entities.
- **Clearly communicating,** in an easily accessible and visible way, as regards **users' inputs and interactions with an AI system**, **any data-processing activities, how data is stored, and whether the data is shared with third parties**. Users should be able to **opt out** of their **input data and interactions being stored and used**, and opt-out should be the **default setting**.

Regulation is needed to impose guardrails on AI companies and entities to build responsible AI systems, mandating them to:

- Provide **information about their training datasets for public scrutiny** in an easily accessible and understandable manner, including a searchable database.
- Conduct **impact assessments to check for bias – including diversity and representation, and inaccuracies and misrepresentation in different languages** – *before* AI systems are deployed, and to review them on an ongoing basis.
- **Implement democratic governance structures**, which can take different forms, such as a **supervisory council, citizen assembly, or employer- and user representation.**
- Conduct systemic **risk assessment, assessing risks to the information space pre-release, and undergo a third-party conformity assessment for medium- and high-risk systems**.

Efforts are needed to strengthen trust in the information space by:

- **Establishing standards governing content authenticity and provenance, including for author authentication**, and using these standards in government communication and media.
- Mandating **platforms to detect authenticity and provenance information and AI-generated content** by the best means currently available and **displaying this information to the end user.**
- Mandating **deployers of synthetic entities** (e.g., chatbots, virtual assistants) **to make users aware that they are engaging with an AI-driven interactive system** and to **support methods for**

reliably detecting the content they generate by, among others, embedding a watermark in generated content.

- Considering the provision of public funding to support the development and maintenance of **public infrastructure for trustworthy AI systems**. This infrastructure would include **public training datasets, public alternatives to for-profit recommender, content moderation or classifier systems, public alternatives to for-profit generative AI systems, and alternative digital information space infrastructure.**
- Promoting the creation of a tailored **certification system for AI companies** inspired by the success of the **Fair Trade certification system**.
- **Adopting codes and guidelines for responsible government and media use of AI systems**, such as the Paris Charter on AI and Journalism, to set clear guidelines on the use of AI.
- Establishing **enforceable rights for media organizations and journalists including a right to know** (about the use of their content in AI systems), **a right to opt out and a right to fair compensation**.

Clear accountability for harms in the information space must be established by:

- Implementing a **fault-based liability regime for AI developers and deployers** regarding the outputs of their systems. Developers and deployers would be held liable for their failure to comply with obligations related to risk mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise. Furthermore, it should be made specifically clear that the **burden of proof lies with AI developers and deployers** in cases where individuals or a group has been damaged.
- Implementing a **strict liability regime for developers and deployers of AI systems utilized to microtarget users based on protected characteristics or special categories of personal data.**
- **Introducing a rebuttable presumption that platforms are liable for illegal content they host and the harm they cause unless they can prove that they have** implemented comprehensive risk-mitigation measures, complied with transparency requirements, and adopted state-of-the-art detection, labeling, and provenance and authenticity standards.
- Clarifying by legal means that **content generated by AI** should **not be considered as third-party content or hosting content** by the AI system in determining the liability of the generative AI deployer.
- **Developing a comprehensive legal framework that clearly defines the rights of individuals** including the right to be informed, to receive an explanation, to challenge an outcome, and to non-discrimination, **and mandating AI systems to establish complaint-handling procedures**.
- **Appointing an AI Ombudsman or strengthening an existing Ombudsman institution** tasked with reviewing unsettled complaints, and to serve as a representative of the plaintiff if no amicable solution can be found.

AI systems need to be independently overseen and scrutinized by:

- **Establishing a new authority or enhancing the capabilities of existing authorities** and entrusting it/them with the mandate to oversee the enforcement of AI regulations and issue implementing acts.
- **Establishing and funding an independent AI research body,** which can be either national or supranational, composed of several independent research laboratories.

- **Committing to financially compensating CSOs participating in official oversight and AI governance institutions and structures.**
- **Mandating AI developers and deployers to provide transparency** about their systems in a **tiered approach, providing information to the general public and more detailed information to regulators and vetted researchers.**
- Mandating platforms **to grant vetted researchers the possibility to conduct experimental evaluations of AI systems and to establish an "accountability sandbox" accessible to external stakeholders.**
- **Establishing a tax on AI companies and entities to address the societal impact of AI.** A portion of the revenue generated by this tax should be allocated to fund community-based AI literacy programs, public alternatives to for-profit systems, and civil society initiatives.
- **Setting up strong legal protections for whistleblowers** who qualify as former or current employees in the AI industry.
- Ensuring that international AI governance is governed by **democratic principles in strengthening cooperation through the Partnership for Information and Democracy, and** in promoting the formation of **a Global AI Forum for Open Dialogue, with sustained and equal participation of civil society, media and journalists, researchers, and other community and public interest organizations.**

ABOUT THE **WORKING GROUP**

The Working Group on Artificial Intelligence and its implications for the information and communication space was launched on 28 September 2023, the International Day for Universal Access to Information. The Working Group is composed of 14 renowned global researchers and experts from diverse academic and applied disciplines, who offered guidance to the rapporteurs and to the Forum in setting the thematic focus of the report and developing its recommendations.

CO-CHAIRS OF THE WORKING GROUP:

- **Laura Schertel Mendes,** Professor of Law at the Brazilian Institute for Development, Education and Research (IDP), and Rapporteur of the Jurists Commission advising the Brazilian Senate on AI regulation, Brazil

- **Jonathan Stray,** Senior Scientist, UC Berkeley Center for Human-Compatible AI, US

MEMBERS OF THE WORKING GROUP:

- **Rachel Adams**, Director, Global Index on Responsible AI and African Observatory on Responsible AI, Research ICT Africa, South Africa

- **Linda Bonyo**, Founding Director, Africa Law Tech; Founder of the Lawyers Hub, Kenya

- **Marta Cantero Gamito**, Professor of IT Law, University of Tartu; Research Fellow, Florence School of Transnational Governance (Chair on AI & Democracy, EUI), Italy

- **Alistair Knott**, Professor of AI, Victoria University of Wellington, New Zealand

- **Syed Nazakat**, Founder and CEO, DataLEADS, India

- **Alice Oh**, Professor, Korea Advanced Institute of Science and Technology, Korea

- **Prof. Alejandro Pisanty**, Faculty of Chemistry, National Autonomous University of Mexico (UNAM), Mexico

- **Gabriela Ramos**, Assistant Director-General for Social and Human Sciences, UNESCO

- **Prof. Dr Achim Rettinger**, Computational Linguistics, Trier University, Germany

- **Prof. Edward Santow**, Co-Director, Human Technology Institute, University of Technology Sydney, Australia

- **Dr Suzanne Vergnolle**, Associate Professor of Technology Law, Cnam, France

- **Claes de Vreese**, Distinguished University Professor of AI and Society with a special focus on Media and Democracy, University of Amsterdam, Netherlands

LEAD RAPPORTEUR:

- **Viviana Padelli,** Economist and Public Policy Professional, working at the intersection of tech and democracy

RAPPORTEURS:

- **Lia Chkhetiani**, Political Scientist and Lawyer
- **Kaye Celine Palisoc**, Computational and Public Policy Analyst

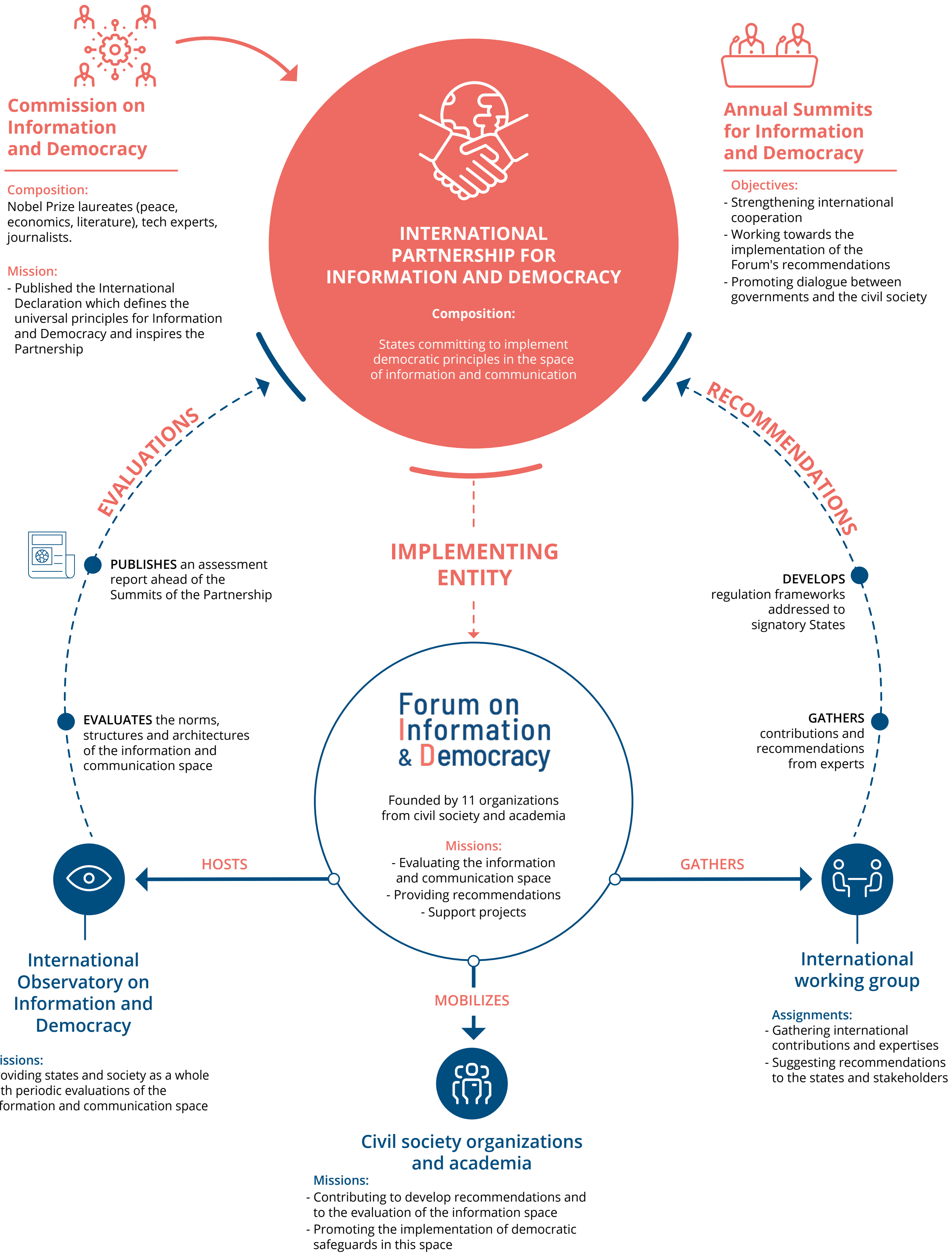
FORUM ON INFORMATION AND DEMOCRACY:

- **Christophe Deloire**, Chair
- **Michael Bąk**, Executive Director
- **Camille Grenier**, Operations Director
- **Katharina Zügel**, Policy Manager
- **Julie Pailhes**, Project Officer

The report of this Working Group reflects the views expressed during the rapporteur teams' discussions with the members, interviews with subject-matter experts, and in written contributions received from experts and organizations engaged in the field. The team of rapporteurs did not seek unanimity on every conclusion or recommendation, recognizing that diverse perspectives could not always be reconciled. This report should not be understood to be the result of a formal negotiation validated by the Working Group members, but as the rapporteur team's best efforts to offer a path forward.

ABOUT **THE FORUM ON INFORMATION AND DEMOCRACY**

Providing democratic safeguards for the global communication and information space



Commission on Information and Democracy

Composition:
Nobel Prize laureates (peace, economics, literature), tech experts, journalists.

Mission:
- Published the International Declaration which defines the universal principles for Information and Democracy and inspires the Partnership

INTERNATIONAL PARTNERSHIP FOR INFORMATION AND DEMOCRACY

Composition:

States committing to implement democratic principles in the space of information and communication

Annual Summits for Information and Democracy

Objectives:
- Strengthening international cooperation
- Working towards the implementation of the Forum's recommendations
- Promoting dialogue between governments and the civil society

EVALUATIONS

RECOMMENDATIONS

PUBLISHES an assessment report ahead of the Summits of the Partnership

EVALUATES the norms, structures and architectures of the information and communication space

DEVELOPS regulation frameworks addressed to signatory States

GATHERS contributions and recommendations from experts

IMPLEMENTING ENTITY

Forum on Information & Democracy

Founded by 11 organizations from civil society and academia

Missions:
- Evaluating the information and communication space
- Providing recommendations
- Support projects

HOSTS

GATHERS

MOBILIZES

International Observatory on Information and Democracy

Missions:
- Providing states and society as a whole with periodic evaluations of the information and communication space

International working group

Assignments:
- Gathering international contributions and expertises
- Suggesting recommendations to the states and stakeholders

Civil society organizations and academia

Missions:
- Contributing to develop recommendations and to the evaluation of the information space
- Promoting the implementation of democratic safeguards in this space

GLOSSARY

A/B testing: compares the performance of two versions of content to see which one appeals more to visitors/viewers. It tests a control (A) version against a variant (B) version to measure which one is most successful based on key metrics[1]

Accountability sandbox: a testing environment that allows external stakeholders to input data and gain insights into an AI system's functioning

Alignment: the process of making AI systems behave in line with human intentions and values[2]

Artificial intelligence: a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments[3]

AI agent: an AI system characterized by its ability to independently perform actions that consistently contribute to achieving complex goals defined by humans over extended periods, with minimal direct human supervision[4]

AI Companies and Entities: companies and any other actors (i.e., non-profits, research institutions, etc.) that develop and/or deploy artificial intelligence systems. AI development and/or deployment can be only one part of their business

AI Developer: designs, codes, or produces AI systems[5]

AI Deployer: decides on the use of the AI-system, exercises control over the associated risk and benefits from its operation[6]

AI Model: a program that has been trained on a set of data to recognize certain patterns or make certain decisions without further human intervention. AI models apply different algorithms to relevant data inputs to achieve the tasks, or output, they have been programmed for[7]

API (application programming interface): mechanisms that enable two software components to communicate with each other using a set of definitions and protocols[8]

AI Subject: any entity (i.e., person, agency, organization, etc.) which is the subject of an AI system (i.e., whose data is used in the training; about whom AI outputs are created)

AI System: AI-based component, software and/or hardware. Usually AI systems are embedded as components of larger systems, rather than stand-alone systems[9]

AI User: an individual or entity utilizing AI systems

Algorithm: a computational process used to make decisions[10]

[1] Oracle.com (2022). *What is A/B Testing?* Available at: www.oracle.com/cx/marketing/what-is-ab-testing/ (Accessed: 2 February 2024)

[2] Ji, J., et al. (2023). *AI Alignment: A Comprehensive Survey*. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2310.19852. (Accessed 2 February 2024)

[3] As defined by OECD *OECD AI Principles overview*. Available at: https://oecd.ai/en/ai-principles (Accessed: February 7 2024).

[4] Shavit, Y. et al. (2023). *Practices for Governing Agentic AI Systems.* Available at: https://openai.com/research/practices-for-governing-agentic-ai-systems (Accessed: 5 February 2024)

[5] BSA (2023). *AI Developers and Deployers: An Important Distinction | BSA | The Software Alliance*. [online] Available at: www.bsa.org/policy-filings/ai-developers-and-deployers-an-important-distinction (Accessed: 2 February 2024).

[6] Law Insider. (n.d.). *deployer Definition*. [online] Available at: www.lawinsider.com/dictionary/deployer (Accessed: 2 February 2024).

[7] IBM (n.d.). *What is an AI model? | IBM*. [online] Available at: www.ibm.com/topics/ai-model (Accessed: 2 February 2024).

[8] Amazon Web Services (2023). *What is an API? - API Beginner's Guide - AWS*. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/what-is/api/ (Accessed: 2 February 2024).

[9] As defined by the European Commission (2018), *A Definition of AI: Main Capabilities and Scientific Disciplines*, p.1. Available at: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (Accessed: 8 February 2024)

[10] As defined by the Forum on Information and Democracy (2023). *Pluralism on News and Information in Curation and Indexing Algorithms*, p. 14. Available at: https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf (Accessed: 8 February 2024)

Bot (social media bot): programs that vary in size depending on their function, capability, and design, and can be used on social media platforms to do various useful or malicious tasks while simulating human behaviors[11]

Civil Society Organization: any non-profit, voluntary citizens' group which is organized on a local, national or international level[12]

Classifiers: an algorithm that automatically orders or categorizes data into one or more of a set of "classes". The process of categorizing or classifying information based on certain characteristics is known as classification[13]

Content Provenance and Authenticity: the facts about the origin and history of a piece of digital content assets (image, video, audio recording, document)[14]

Conformity Assessment: an ex-ante form of human oversight, which aims to verify that AI systems comply with recognized technical, ethical, and legal standards[15]

Criminal liability: a legal standard according to which individuals are held responsible for their actions if they have committed a criminal act[16]

Curation: decisions related to filtering and ranking content for human attention[17]

Data Augmentation: a technique of artificially increasing the training set by creating modified copies of a dataset using existing data. It includes making minor changes to the dataset or using deep learning to generate new data points[18]

Data Curation: process of organizing, describing, cleaning, enhancing, and preserving data for public use[19]

Data Provenance: documentation about where a piece of data comes from and how it came to be in its current state[20]

Data Segments: subset of a dataset, usually grouped according to similar attributes

Disinformation: information that is false and deliberately created to harm a person, social group, organization or country[21]

Fault-based liability: in civil law, fault-based liability is a legal standard according to which a defendant is only held responsible for the consequences of his actions if his fault (intent or negligence) has been proven. In practice, this means that the plaintiff must prove that the damage was caused by the defendant's misconduct[22]

[11] As defined by the US Office of Cyber and Infrastructure Analysis (2018). Social Media Bots Overview. Available at: https://niccs.cisa.gov/sites/default/files/documents/pdf/ncsam_socialmediabotsoverview_508.pdf?trackDocs=ncsam_socialmediabotsoverview_508.pdf (Accessed: 8 Feb. 2024).

[12] United Nations (n.d.). *Civil society*. [online] United Nations. Available at: www.un.org/en/civil-society/page/about-us (Accessed: 8 Feb. 2024).

[13] DeepAI. (2019). *Classifier*. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/classifier (Accessed: 2 Feb 2024).

[14] c2pa.org. (n.d.). *FAQ - C2PA*. [online] Available at: https://c2pa.org/faq/ (Accessed: 31 January 2024).

[15] Mökander, J., Axente, M., Casolari, F. and Floridi, L. (2021). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, 32. doi:https://doi.org/10.1007/s11023-021-09577-4.

[16] Sneha Solanki (2024). *What is criminal liability? Definition and resources for defense attorneys*. [online] Thomson Reuters Law Blog. Available at: https://legal.thomsonreuters.com/blog/what-is-criminal-liability/ (Accessed: 9 Feb. 2024).

[17] As defined by Forum on Information and Democracy (2023). *Pluralism on News and Information in Curation and Indexing Algorithms*, p 14. Available at: https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf (Accessed: 8 February 2024).

[18] Awan, A.A. (2022). *A Complete Guide to Data Augmentation*. [online] www.datacamp.com. Available at: www.datacamp.com/tutorial/complete-guide-data-augmentation (Accessed: 31 January 2024).

[19] ICPSR (n.d.). *Data Management & Curation*. [online] Available at: www.icpsr.umich.edu/web/pages/datamanagement/index.html (Accessed: 8 February 2024).

[20] faculty.washington.edu. (n.d.). *About Data Provenance*. [online] Available at: https://faculty.washington.edu/hazeline/ProvEco/generic.html (Accessed: 1 February 2024).

[21] As defined by UNESCO (2020). Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training. Available at https://en.unesco.org/fightfakenews (Accessed 8 February 2024).

[22] Coleman, J.L. (2002). Fault and strict liability. *Risks and Wrongs*, pp.212–233. doi: https://doi.org/10.1093/acprof:oso/9780199253616.003.0012 (Accessed: 8 February 2024).

Fine-tuning: the process of adjusting the parameters of a pre-trained large language model to a specific task or domain. Fine-tuning addresses the lack of model specialization in some areas by allowing the model to learn from domain-specific data to make it more accurate and effective for targeted applications[23]

Generative AI: Generative AI refers to a category of AI algorithms that generate new outputs based on the data they have been trained on[24]

Human Labeling for AI training: involving human inputs in the process of labeling training data[25]

Interoperability: the ability to transfer and render useful data and other information across systems, applications, or components, including platforms[26]

Large Language Models (LLM): a type of artificial intelligence model that has been trained through deep-learning algorithms to recognize, generate, translate, and/or summarize vast quantities of written human language and textual data[27]

Licencing: an ex-ante tool of human oversight, which entails a comprehensive assessment of the AI system against legal, ethical, and technical requirements, after which the AI systems are granted approval to be placed on the market[28]

Media: those responsible for the periodic creation of information and content and its dissemination, over which there is editorial responsibility, irrespective of the means and technology used for delivery, which are intended for reception by, and which could have a clear impact on, a significant portion of the general public[29]

Misinformation: information that is false but not created with the intention of causing harm[30]

Model Cards: documentation of AI uses and limitations, mitigation strategies, and descriptions of planned training and testing procedures – prior to and during system development

Moderation: when online platforms screen and monitor user-generated content based on platform-specific rules and guidelines to determine whether to host or continue hosting a specific piece of content under their terms of service. These decisions include removal of content, permanently, temporarily, or by geographical area[31]

Natural Language Processing: a set of methods for making human language accessible to computers[32]

Open Source: software published under an open source license which offers the freedoms to use, study, share and improve the software[33]

[23] Turing (n.d.). *Fine-Tuning LLMs: Overview, Methods & Best Practices*. [online] Available at: www.turing.com/resources/finetuning-large-language-models#what-is-fine-tuning (Accessed 9 February 2024).

[24] Routley, N. (2023). *What is generative AI? An AI explains*. [online] World Economic Forum. Available at: www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/ (Accessed: 9 February 2024).

[25] Mehta, R. (2023). *Human Data Labeling for Successful AI*. [online] iMerit. Available at: https://imerit.net/blog/human-data-labeling-for-successful-ai/ (Accessed: 9 February 2024).

[26] As defined by Forum on Information and Democracy (2023). *Pluralism on News and Information in Curation and Indexing Algorithms*, p.14. Available at: https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf (Accessed: 8 February 2024).

[27] Sarno, I. (2023). *What Is a Large Language Model?* [online] knowledge-centre-interpretation.education.ec.europa.eu. Available at: https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model (Accessed 8 February 2024).

[28] Malgieri, G. and Pasquale, F. (2024). Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review*, [online] 52, p.105899. doi:https://doi.org/10.1016/j.clsr.2023.105899.

[29] As defined by Forum on Information and Democracy (2023). *Pluralism on News and Information in Curation and Indexing Algorithms*, p. 14. Available at: https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf (Accessed: 8 February 2024).

[30] As defined by Forum on Information and Democracy (2020). *How to End Infodemics*, p. 16. Available at: https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pd

[31] As defined by Forum on Information and Democracy (2022). *Accountability Regimes for Social Networks and their Users*, p.10. Available at: https://informationdemocracy.org/wp-content/uploads/2023/01/ID_Report-on-Accountability-regime_Sept22.pdf (Accessed: 8 February 2024).

[32] Eisenstein, J. (2019). *Introduction to natural language processing*. Cambridge, Massachusetts: The Mit Press.

[33] FSFE - Free Software Foundation Europe (n.d.). *What is Free Software*. Available at: https://fsfe.org/freesoftware/freesoftware.en.html (Accessed: 7 February 2024).

Platform: entities that help structure the information and communication space by creating the technical means, architecture, and standards for information and communication[34]

Public alternative to for-profit system: a system developed and run in the public interest by an organization or several organizations acting in the public interest. This organization can take various forms, such as a research or civil society organization, an independent administrative authority, or a status similar to a public service media, among others

Recommender system: systems that algorithmically suggest items to a user, potentially based on information about the user (profiling them based on their interests); information about the content (including signals such as indexing and/or predicting virality); and/or the organizations' interests[35]

Red-Teaming: process of emulating a potential adversary's attack or exploitation capabilities against an enterprise's security posture, carried out by a group of authorized people called a "red team"[36]

Strict liability: in civil law, strict liability is a legal standard under which a defendant is held responsible for the consequences of his or her actions regardless of their fault (intent or negligence). In practice, this means that fault is not a necessary factor in determining liability[37]

Synthetic Data: information that has been generated on a computer to augment or replace real data to improve AI models, protect sensitive data, and mitigate bias[38]

Synthetic Entity: artificial constructs, initiated digital or physical forms, engineered to imitate or reproduce specific characteristics of natural entities (i.e., chatbots, virtual assistants)[39]

Training Data: information/data used during the model learning process

Troll: an entity who intentionally disrupts online communities[40]

Watermarking: the process of embedding an identifying pattern in a piece of content in order to track its origin[41]

[34] For a more detailed explanation see Forum on Information and Democracy (2020). *How to End Infodemics*, p. 16. Available at: https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf (Accessed: 8 February 2024).

[35] As defined by Forum on Information and Democracy (2023). *Pluralism on News and Information in Curation and Indexing Algorithms*, p. 14. Available at: https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf (Accessed: 8 February 2024).

[36] NIST (n.d.). *Computer Security Resource Center - Glossary*. [online] csrc.nist.gov. Available at: https://csrc.nist.gov/glossary/term/Red_Team (Accessed: 8 February 2024).

[37] J.L. (2002). Fault and strict liability. *Risks and Wrongs*, op. Cit

[38] Martineau, K. (2021). *What is synthetic data?* [online] IBM Research Blog. Available at: https://research.ibm.com/blog/what-is-synthetic-data (Accessed: 31 January 2024).

[39] Nanni, D. (2023). *Synthetic Entities: Definitions, Characteristics, and Future Perspectives*. [online] Brass For Brain. Available at: https://medium.com/brass-for-brain/synthetic-entities-definitions-characteristics-and-future-perspectives-49673f22f6fe (Accessed: 31 January 2024).

[40] Schwartz, M. (2008). The Trolls Among Us. *The New York Times*. [online] 3 Aug. Available at: https://www.nytimes.com/2008/08/03/magazine/03trolls-t.html (Accessed: 7 February 2024).

[41] Brookings. (n.d.). *Detecting AI fingerprints: A guide to watermarking and beyond*. [online] Available at: www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/ (Accessed 31 Jan. 2024).

INTRODUCTION

The increasingly advanced capabilities and commercial proliferation of Artificial Intelligence (AI) systems such as ChatGPT, “smart bots”, and computationally charged propaganda, underscore their ubiquity, sophistication, and increasing potential to profoundly impact democratic processes, including the information and communication space. Even though AI systems can also be leveraged for better democratic outcomes, they pose considerable challenges and unresolved questions that require wide attention.[42]

For this reason, regulating the development, deployment, and use of AI is high on the political agenda. Examples of the efforts to regulate AI systems include President Biden’s AI Executive Order in the United States,[43] the signing of the Bletchley Declaration by more than 20 countries,[44] the *Hiroshima Process International Guiding Principles for Advanced AI* system issued by the G7 countries in 2023,[45] the adoption of the UNESCO *Recommendations on the Ethics of Artificial Intelligence* in 2021,[46] and of the OECD *Principles on Artificial Intelligence* in 2019,[47] the ongoing negotiations for a Council of Europe Framework Convention on AI,[48] an EU Artificial Intelligence Act,[49] Brazil’s Draft Law Number 2338 on the Use of AI,[50] and Canada’s Draft Artificial Intelligence and Data Act.[51]

The imperative to address the threats posed by AI to the information and communication ecosystem is exceptionally pressing in 2024, as three billion people worldwide will vote in major elections.[52] The implications for democratic processes of newly emerged generative AI systems for microtargeting, and for the production of personifications of candidates to manipulate voters, remain unknown.[53]

[42] Kreps, S. and Kriner, D. (2023). *How AI Threatens Democracy. Journal of Democracy*. Available at: www.journalofdemocracy.org/articles/how-ai-threatens-democracy/ (Accessed: 7 February 2024).

[43] The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* Available at: www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Accessed: 7 February 2024).

[44] *The Bletchley Declaration by Countries Attending the AI Safety Summit.* 1-2 November 2023. Available at: www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 (Accessed: 7 February 2024).

[45] European Commission (2023). *Hiroshima Process International Guiding Principles for Advanced AI systems*. available at: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system (Accessed: 7 February 2024).

[46] UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence.* Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

[47] OECD (2019). *Recommendation of the Council on Artificial Intelligence.* Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed: 7 February 2024).

[48] Council of Europe (2023). *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law.* Available at: https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043 (Accessed: 7 February 2024).

[49] On February 2, the Council of the EU countries approved the final version of the EU AI Act, the draft of which was also leaked online. On February 13, the Internal Market and the Civil Liberties, Justice, and Home Affairs (LIBE) Committees of the European Parliament endorsed a provisional agreement on the AI Act, paving the way for the final vote by the legislative assembly scheduled for April 2024.
Bertuzzi, L. (2024). *EU countries give crucial nod to first-of-a-kind Artificial Intelligence law.* Euractiv. European Council (2024). *Artificial Intelligence Act: Council and Parliament strike a deal on the first rules for AI in the world.* Available at: www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ (Accessed: 7 February 2024). Available at: www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/ (Accessed: 7 February 2024). The latest version of the draft EU AI Act accessed on 7 February 2024. Available at: https://drive.google.com/file/d/13qcPGQNFHTcfg4XxIybibFIDnkWTHBbu/view?_hsmi=292273120&*hsenc=p2ANqtz-83laGl4OZccMJdsws8KrlMygORyZSiS1yD7lHUY8dbIDZzVdNvqPoBwq5n2V3GczVGWk1nWLJyuDlHeodJ3HJxNK4ia403C9b8JBLyBxh4CZ7iPHc* (Accessed: 7 February 2024).
Yun Chee, F (2024). EU lawmakers ratify political deal on artificial intelligence rules. Reuters. Available at: https://www.reuters.com/technology/eu-lawmakers-back-political-deal-artificial-intelligence-rules-2024-02-13/ (Accessed: 13 February 2024).

[50] The National Congress of Brazil (2023). *Bill 2338/2023 ‘Brazilian Artificial Intelligence Act’*. Available at: https://mcusercontent.com/af97527c75cf28e5d17467eaa/files/248d109f-eeef-7496-4df1-12d29affb522/PL_23382023_Senado_ENG_VF.pdf (Accessed: 7 February 2024).

[51] Government of Canada (2023). *Artificial Intelligence and Data Act*. Available at: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act (Accessed: 7 February 2024).

[52] Hsu, T., Thompson, S.A. and Myers, S.L. (2024). *Elections and Disinformation Are Colliding Like Never Before in 2024.* The New York Times. 9 Jan. Available at: www.nytimes.com/2024/01/09/business/media/election-disinformation-2024.html (Accessed: 7 February 2024).

[53] *ibid*.

The urgency of the issue is also recognized in the *Global Risks Report 2024* recently published by the World Economic Forum, which considers AI-powered and social-media-driven misinformation and disinformation as the world's biggest short-term threats, especially in the context of their negative impact on the credibility of elections, trust in information, and political polarization.[54]

The most pressing challenges related to the influence of AI on the information and communication space require broad, urgent, and multi-stakeholder action defined by democratic principles and processes. While States and international organizations are taking action within their means, most of these initiatives do not necessarily focus on the information and communication space and often convey general principles rather than concrete policy actions.

The Forum on Information and Democracy takes a holistic approach to safeguarding democratic rules in the information and communication space by proposing practicable and comprehensive policy recommendations to ensure that AI contributes to information integrity, rather than being a risk.

AI developers and deployers are at the forefront of building trustworthy AI systems working in the public interest. Therefore, Chapter 1 defines guardrails for the development and deployment of AI systems impacting the information and communication space. In addition, liability and accountability regimes are proposed to uphold not only individual but also collective rights.[55] To this end, Chapter 2 defines accountability and liability regimes for the developers, deployers and users within the information and communication ecosystem. In addition to being reactive (i.e., enforcing compliance and providing mechanisms for redress), accountability regimes should also support proactive approaches to incentivize an environment conducive to ethical AI. Consequently, Chapter 3 proposes incentive schemes to encourage ethical development, deployment, and use of AI systems. Finally, in addition to governance frameworks at national and international levels, stakeholder involvement, scrutiny by researchers, and robust processes for the public release of AI systems are essential to enforce democratic rules of AI. Thus, Chapter 4 proposes holistic governance options and enforcement of rules for regulating and governing AI systems.

AI and Information Creation

AI systems, particularly generative AI systems, are revolutionizing the way we create information across various mediums, including text, audio, images, and video, presenting both challenges and opportunities:

> Generative AI systems are being developed and used in ways that may not respect copyright, data protection, or privacy laws. These include using copyrighted material without proper attribution or inadvertently leaking private or sensitive information from their training data, which can be a concern for freedom of expression and privacy.[56]

> Generative AI raises challenges to the authenticity and trustworthiness of the content produced. AI can invent sources and create misinformation (hallucinations) and deepfakes. That information can be used unintentionally or intentionally to spread mis- and disinformation, create chaos, deceive, erode trust, and jeopardize public discourse. Moreover, generative AI systems lower the barriers for producing deceptive content, enabling a larger and more diverse group of

[54] The World Economic Forum (2024). *The Global Risks Report 2024*. Available at: www.weforum.org/publications/global-risks-report-2024/ (Accessed: 7 February 2024). Associated Press (2024). *AI-powered misinformation is the world's biggest short-term threat, Davos report says.* Available at https://apnews.com/article/artificial-intelligence-davos-misinformation-disinformation-climate-change-106a1347ca9f987bf71da1f86a141968 (Accessed: 7 February 2024).

[55] The concept of protecting collective rights is very common in Brazil. In this sense, the draft AI Law (Nr 2338, 2023) includes this aspect in Article 6: The defense of the interests and rights provided for in this Law may be exercised before the competent administrative bodies, as well as in court, individually or collectively, in accordance with the provisions of the relevant legislation regarding individual, collective and diffuse protection instruments. The National Congress of Brazil (2023). *Bill 2338/2023 'Brazilian Artificial Intelligence Act'*. Available at: https://mcusercontent.com/af97527c75cf28e5d17467eaa/files/248d109f-eeef-7496-4df1-12d29affb522/PL_23382023_Senado_ENG_VF.pdf (Accessed: 7 February 2024).

[56] Allen, D. and Weyl, E.G. (2024). The Real Dangers of Generative AI. *Journal of Democracy*. Available at: www.journalofdemocracy.org/articles/the-real-dangers-of-generative-ai/ (Accessed: 7 February 2024).

propagandists to create and spread more convincing, diverse, and tailored disinformation and propaganda at scale.[57]

- AI-generated content often reflects biases against historically marginalized and minoritized groups, influenced by the data on which AI systems are trained. To avoid AI systems exacerbating inequalities, various precautions should be taken into account when developing them. In particular, if training datasets are made more representative than the prevailing realities, AI has the potential to amplify the voices of underrepresented groups.
- While generative AI has the potential to improve information sorting, organization, and customization, its use by journalists and media organizations without proper human oversight can lead to the dissemination of biased or unverified news, ultimately eroding trust in media.

AI and Information Dissemination

AI systems, including those deployed for content moderation, recommendation, ranking, and generative AI, play a significant role in the dissemination of information online. Their use comes with profound implications:

- Generative AI can be employed to create social media bots, amplify mis- and disinformation, and give it credibility, thus contributing to information chaos. This technology can also be used to proliferate armies of AI "trolls" active on social media and in astroturfing activities,[58] dramatically increasing the scale and reach of disinformation and propaganda operations.[59]
- AI systems are tasked with crucial decisions in the information and communication space, as the sheer amount of information available and content created exceeds human capacities to consume, sort, moderate, and verify this information. AI recommender systems decide which information we see, significantly influencing the information landscape and possibly creating echo chambers and filter bubbles.[60]
- Ranking and moderating AI systems suffer from biases, their decision-making processes are not transparent, and they are often optimized for engagement without considering the sociopolitical risks of the content and actors they amplify or analyze. AI systems moderate content relying on their own secret training datasets and definitions, while the moderation policies, accuracy statistics, or transparency reports are generally not published.
- The influence of AI systems and thus of their developers and deployers, on the information and communication space is growing, reinforcing the power of private tech corporations over democratic processes.
- AI systems can also be abused by States. Governments can instruct AI entities to use AI systems to remove unwanted content, thereby strengthening censorship. In general, users have little control or agency when interacting with AI recommender and moderation systems.
- Even though AI-generated content is widely disseminated, online platforms hosting it do not have comprehensive policies related to AI-generated content. Moreover, it is often ambiguous who should be subject to civil liability for the harm caused by the AI-generated content disseminated through social media platforms, significantly increasing the risk for victims seeking compensation.[61]

[57] Goldstein, J. et al (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. Available at: https://arxiv.org/pdf/2301.04246.pdf (Accessed: 7 February 2024).

[58] "Digital astroturfing is a form of manufactured, deceptive, and strategic top-down activity on the Internet initiated by political actors that mimics bottom-up activity by autonomous individuals". Kovic, M et al. (2018). *Digital astroturfing in politics: Definition, typology, and countermeasures. Studies in Communication Sciences*, 18(1). Available at: www.hope.uzh.ch/scoms/article/view/j.scoms.2018.01.005/991 (Accessed: 7 February 2024).

[59] Zhou, J et al (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.* Available at: https://jiaweizhou.me/assets/chi23_ai_misinfo.pdf (Accessed: 7 February 2024).

[60] Arguedas, A. et al (2022). *Echo chambers, filter bubbles, and polarisation: a literature review. Reuters Institute for the Study of Journalism.* Available at: https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review (Accessed: 7 February 2024).

[61] Karathanasis, A.L et al (2022). *Civil Liability for AI Systems: Comment on EU Commission's Proposals*. MIAI. Available at: https://ai-regulation.com/eu-commission-proposals-on-ai-civil-liability/ (Accessed: 7 February 2024).

- Digital exploitation of personal information and microtargeted advertising can be intensified through sophisticated ranking and content moderation algorithms as well as through generative AI, which decreases personalization efforts thanks to "personality inference" from texts.[62]
- Users do not know what information they are automatically excluded from or explicitly exposed to. Overtly personalized messages can be used to exploit their fears and manipulate behavior which can harm democracy.[63]

AI and Information Consumption

AI has the potential to change how we perceive the information we consume:

- It is becoming increasingly difficult to distinguish AI-generated and human-created content, diminishing trust in the global information space. According to some studies, AI misinformation is harder to detect as it meets criteria of credibility, transparency, and comprehensiveness, thus propaganda or disinformation is less discoverable.[64] Furthermore, watermarks can be removed or altered, making it even more difficult to reliably detect the synthetic nature of the content. Even when detection mechanisms are used to verify content, these can be wrong, leaving no dependable method to unequivocally determine whether content is AI-generated.[65]
- AI systems have the capability to provide users with information more tailored to their interests. Yet, this hyper-personalization can also lead to a fragmented information landscape and hinder our ability to recognize trustworthy information. The personalization of targeted information limits access to pluralistic information, exacerbates political polarization, and hinders the construction of public discourse.

Reflecting on the challenges and limitations described above, the following chapters provide policy-oriented and practicable recommendations aimed at helping policymakers create comprehensive, nuanced, and democratic frameworks that harness the innovative potential of AI systems while effectively managing the associated risks to the information and communication space.

[62] Simchon, A. et al. (2024). *The persuasive effects of political microtargeting in the age of generative AI.* Available at: https://academic.oup.com/pnasnexus/article/3/2/pgae035/7591134 (Accessed: 7 February 2024).

[63] Wheeler, T. (2023). *The three challenges of AI regulation*. Brookings. Available at: www.brookings.edu/articles/the-three-challenges-of-ai-regulation/ (Accessed: 7 February 2024).

[64] Zhou, J et al (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.* Available at: https://jiaweizhou.me/assets/chi23_ai_misinfo.pdf (Accessed: 7 February 2024).

[65] Leibowicz, C. (2023). *Why watermarking AI-generated content won't guarantee trust online.* MIT Technology Review. Available at: www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/ (Accessed: 7 February 2024).

CHAPTER 1: DEVELOPING AND DEPLOYING SAFE AND RESPONSIBLE AI SYSTEMS IN THE INFORMATION AND COMMUNICATION SPACE

PART 1

INTRODUCTION

Artificial Intelligence (AI) systems are fundamentally reshaping how we create, access and interact with information. They play a critical function in shaping public opinion, influencing the democratic discourse, and defining the future of journalism. AI-driven algorithms serve pivotal roles in content moderation and curation, and are instrumental to surveillance-based, hyper-targeted advertising. Importantly, the widespread availability of user-friendly interfaces for generative AI systems significantly influences the ease and speed at which the creation of text, audio, imagery, and synthetic data can occur – produced by both good and malicious actors.

The AI industry has become dominated by a small number of mainly for-profit companies in the Global North. Yet, the significant influence on the information space wielded by these private interests remains inadequately overseen by our democratic institutions, which must necessarily provide the required safety protocols, ethical standards, and regulatory frameworks to ensure AI works for all citizens. This gap in democratic oversight allows companies to prioritize market dominance, growth and shareholder value over safety, ethics, and impacts on our institutions, societies, and the architecture of the information space.

The premature deployment of AI risks numerous adverse implications on citizens, our societies, and democratic institutions through the way we create, disseminate and consume information:

- *The hyper-personalization of content through advanced algorithms, fueled by the exploitation of surveillance data and the speed of algorithmic amplification, can lead to the creation of more effective and manipulative mis- and disinformation, resulting in the manipulation of public opinion, user segregation within echo-chambers, and the erosion of diverse, pluralistic information sources.*
- *Hate speech – both illegal and legal-but-still-harmful – enables incitement to violence, strengthens polarization, and harms inter-group relations.*
- *The rise of generative AI makes it increasingly difficult, if not impossible, to discern the authenticity of content, eroding end-users' capacity to verify the provenance and accuracy of information, and ultimately diminishing trust in the information space.*
- *AI's propensity to 'hallucinate' – generating false or misleading information – exacerbates the spread of misinformation, further challenging the reliability and credibility of online content.*
- *Reduced costs and skill requirements for creating disinformation, including deep fakes, lower the barriers for bad actors.*
- *Biases in AI algorithms can distort content creation, dissemination, and decision-making, perpetuating prejudices, reinforcing stereotypes and strengthening inequality.*
- *Heightened risks of violating copyright, privacy, and anti-defamation laws jeopardize intellectual property, personal data and reputation.*

At the same time, AI presents significant opportunities for the democratization of the information space and to enhance the reliability of the information ecosystem. Many of these opportunities – which generally fall short of the priorities for revenue maximization, shareholder value and market monopolization – remain generally untapped.

Design and engineering teams within a small cohort of AI companies and entities building and improving the systems have the unique opportunity to shape the future of AI at the most *fundamental level throughout the design, development and deployment stages. By adopting responsible approaches from the onset and ensuring their teams reflect the diversity of the world in which their products will operate, they have the power to significantly reduce risks and prevent negative consequences in the information and communication space. This requires placing human-centric values, ethics, and diversity at the core of AI development and deployment, integrating these principles throughout the AI lifecycle, and implementing proactive measures to mitigate potential adverse impacts. This also means approaching AI as a public good and not purely as an innovation to generate profits for shareholders.*

A comprehensive approach for assessing AI systems remains a crucial priority. This involves red-teaming exercises to identify and address vulnerabilities to potential harms including, but not limited to, misinformation, manipulation, and malicious use. Only rigorous risk assessments can evaluate AI's biases and privacy implications, ensuring adherence to ethical norms and international legal standards, including international human rights laws and international humanitarian law.[66] *Lastly, post-market monitoring is essential to adapt AI systems to changing information landscapes, ensuring global representations of diversity, and to continuously update AI systems to counter emerging risks.*

1. BUILDING AI SYSTEMS FOR THE SAFE AND INCLUSIVE CREATION, DISTRIBUTION, AND CONSUMPTION OF INFORMATION

Those developing and deploying AI systems impact the information and communication space. They carry profound responsibilities, especially regarding impact on fundamental human rights such as the right to privacy, freedom of opinion and expression, equality and non-discrimination, access to information, and freedom of the press. The Resolution of the UN General Assembly on the *Promotion and Protection of Human Rights in the Context of Digital Technologies*, adopted in 2023, states "that human rights and fundamental freedoms must be respected, protected and promoted throughout the life cycle of artificial intelligence systems and that new and emerging digital technologies should provide new means to promote, protect and exercise human rights and not to infringe upon them".[67]

Design and engineering teams must serve as a crucial first line of defense in upholding human rights, and the right to reliable information as defined by the Partnership for Information and Democracy.[68] Designers and engineers must build AI systems serving democratic values and the public interest.

[66] From this point forward, both international human rights laws and international humanitarian law will be collectively referred to under the term "international human rights laws."

[67] United Nations General Assembly (2023). *Promotion and Protection of Human Rights in the Context of Digital Technologies*, Article 19. Available at: https://digitallibrary.un.org/record/4032837?ln=fr (Accessed: 7 February 2024).

[68] Forum on Information and Democracy. *International Partnership for Information & Democracy*. Available at: https://informationdemocracy.org/international-partnership-on-information-democracy/ (Accessed: 8 February 2024)

This requires a nuanced understanding of AI systems' risks to a plural, free and trustworthy global information space, and a commitment to ongoing learning and adaptation.

In this complex landscape, implementing technical solutions to address the risk of bias must be a top priority. Companies train AI systems on historical data. These risks perpetuating prejudices and violating the right to non-discrimination based on race, color, sex, sexual orientation, gender identity, language, religion, political or other opinion, national or social origin, property, birth, or other status.[69]

Countering cultural hegemony in training data is equally important. AI systems should not only mirror the perspectives from major tech hubs but also represent viewpoints and languages of various cultures and communities of the Global Majority. This requires embedding the experiences, values, and agency of the wider global community into the training of AI systems. Inclusivity in AI not only ensures its global relevance but also supports multilingualism and equal access to knowledge. This approach is vital for preserving cultural diversity and ensuring equity across diverse cultures as they are reflected within these AI systems.[70]

Embedding privacy and intellectual property considerations from the design stage through the AI lifecycle is also crucial. With data as a key asset, maintaining transparency in data use, ensuring robust security, and respecting intellectual property laws and data protection laws are all fundamental.

Developers must prioritize safeguarding AI against misuse and sophisticated adversarial attacks. They must enforce stringent measures for verifying the provenance and authenticity of data – key to enhancing the transparency of AI systems and ensuring their resistance to disinformation. Moreover, for advanced threats, developers must put in place mitigation strategies. These strategies should include measures maximizing the difficulty and expense of any attempts at manipulating AI in ways that could infringe upon human rights.

A variety of international actors have developed ethical principles to guide AI developers and deployers in the design, development and deployment of these systems. These include, among others, UNESCO's *Recommendation on the Ethics of Artificial Intelligence*,[71] which includes a policy chapter specifically focused on communication and information, providing concrete, domain-specific policy recommendations; the *Asilomar AI Principles*;[72] the *Windhoek Statement on Artificial Intelligence in Southern Africa*;[73] the Council of Europe's *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*;[74] the OECD's *Recommendation of the Council on Artificial Intelligence*;[75] the *Hiroshima Process International Guiding Principles for Advanced AI Systems*;[76] the *ASEAN Guide on AI Governance and Ethics*;[77] and the *Ethics Guidelines for Trustworthy Artificial Intelligence* elaborated by the EU High-Level Expert

[69] United Nations. *Universal Declaration of Human Rights,* Article 2. Available at: www.un.org/en/about-us/universal-declaration-of-human-rights (Accessed: 7 February 2024).

[70] United Nations. *Universal Declaration on Cultural Diversity,* Article 6. Available at: https://www.ohchr.org/en/instruments-mechanisms/instruments/universal-declaration-cultural-diversity#:~:text=Freedom%20of%20expression%2C%20media%20pluralism,the%20guarantees%20of%20cultural%20diversity (Accessed: 7 February 2024).

[71] UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

[72] Future of Life (2017). *Asilomar AI Principles*,. Available at: https://futureoflife.org/open-letter/ai-principles/ (Accessed: 7 February 2024).

[73] UNESCO (2022). *Windhoek Statement on Artificial Intelligence in Southern Africa*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000383197 (Accessed: 7 February 2024).

[74] Council of Europe (2023). *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Available at: https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043 (Accessed: 7 February 2024).

[75] OECD (2019). *Recommendation of the Council on Artificial Intelligence*. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed: 7 February 2024).

[76] European Commission (2023). *Hiroshima Process International Guiding Principles for Advanced AI system*. Available at: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system (Accessed: 7 February 2024).

[77] ASEAN (2023). *ASEAN Guide on AI Governance and Ethics*. Available at: https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf (Accessed: 7 February 2024).

Group on AI.[78] Country specific ethical principles could also be developed to reflect social and legal values, provided they align with international human rights law, norms, and standards.

RECOMMENDATIONS
TO **STATES**

> **Offer practical assistance to AI developers and deployers in turning existing international human rights law and AI ethical principles[79] into practicable measures that promote and protect human rights.** This entails fostering compliance by providing detailed guidance and concrete examples, illustrating how to successfully implement existing ethical design principles in practice.

RECOMMENDATIONS
TO **AI COMPANIES AND ENTITIES**[80]

> **Integrate considerations of broad societal impact into the decision-making processes for the design, development, and deployment of AI systems in the information and communication space.** This integration, which could be facilitated by tools such as UNESCO's *Ethical Impact Assessment*,[81] is essential to balance the prevalent emphasis on scaling up, profit maximization, and gaining competitive advantages.

1.1 TRAINING DATASETS

The integrity and quality of any AI system relies fundamentally on the data it learns from, and on those who identify and provide the data. The potential introduction of bias remains a pressing concern during the development process of an AI system, especially during data collection and pre-processing. Bias at this stage can emerge in various forms, such as labeling bias, sampling bias, and exclusion bias arising from the inappropriate removal or omission of relevant data from the dataset.[82] Bias is further deepened when developer teams training these systems lack diversity – *inter alia* cognitive, religious, ethnic, gender, sexual, geographic, and economic diversity. Such biases in the definition and modeling of the training dataset can have profound implications.

First, bias introduced into AI systems by incorporating training data that is unrepresentative can result in the further marginalization of disenfranchised communities. For example, in content moderation on social media platforms, if the training data primarily consists of content flagged as inappropriate by users from specific cultural backgrounds, the AI may develop a skewed understanding of what constitutes offensive content. This skewed understanding could lead to the over-censorship of topics or expressions relevant to underrepresented groups, while failing to identify and address content that

[78] European Commission (2019). Ethics guidelines for trustworthy AI. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed: 15 February 2024).

[79] Globally recognized AI ethical principles include the UNESCO's Recommendation on the Ethics of Artificial Intelligence, as well as other ethical principles listed in the introduction.

[80] Throughout the report the terminology AI company and entity is used to refer to any actor that develops and deploys AI systems even if this is not their main area of activity.

[81] UNESCO (2023). *Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence*. Available at: www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence (Accessed: 7 February 2024).

[82] European Parliament (2022). *Auditing the quality of datasets used in algorithmic decision-making systems*. Available at: www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf (Accessed: 7 February 2024).

is harmful to these communities. To address this effect, it is vital to include a wide range of cultural expressions and nuances within the same language in training datasets. At a minimum, this requires ensuring diverse perspectives, experiences, and backgrounds among the people on relevant teams.

Additionally, bias in AI systems can arise from the inherent limitations in natural language processing tools, which are often developed for one language and may not work as efficiently when applied to other dialects or languages. This issue is particularly acute for languages with smaller online footprints and limited available training data, or training data that only reflects some parts of society or culture. The scarcity of data for these languages limits the algorithms' ability to learn effectively, leading to automated tools with higher error rates.[83] This situation inadvertently reinforces cultural hegemony from dominant regions and languages. Addressing this data scarcity is essential to bridge the digital divide across different global regions, to create AI systems tailored to the needs of the Global Majority and smaller language communities, and to redistribute technological power in a more equitable manner. While multilingual language models show promise, they also have important limitations. Among those are the reliance on machine-translated text that often contains errors and the lack of contextual references.[84] Deploying models prematurely and without proper safeguards can jeopardize freedom of expression for certain geographies or groups by leading to the inaccurate filtering and flagging of specific information, as seen in content moderation systems. Some non-English-speaking democracies face heightened risk due to limited data availability, and in countries where free expression remains under threat or is seriously constrained, by authoritarian or weak democratic governance regimes.[85] Importantly, this risk significantly increases in Global Majority countries and increases even more in those regions with less-commonly spoken languages and dialects.

Furthermore, as AI models become more and more integrated into various aspects of our lives, the challenge of protecting individual privacy escalates. The increasing adoption and continuous evolution of AI necessitate the collection of larger volumes of data (this important aspect will be further discussed in subsections a., b. and e.), potentially leading to more intrusive data gathering methods, including enhanced surveillance techniques and manipulative user interface designs. Such practices amplify concerns about the use and security of personal information within AI systems, posing a significant risk to individual privacy. In particular, if generative AI systems are trained on datasets containing sensitive personal data, including personal identifiers, sexual orientation, or political affiliations (among others), there is a risk they might generate outputs that unintentionally expose this sensitive information.

This issue is compounded by the rapid pace of technological advancements in AI that can challenge and often outpace the current data protection legal frameworks where they exist.

Lastly, it is unclear whether current data collection techniques comply with prevailing legal approaches to intellectual property. Non-compliance in this domain poses risks not only to individual rights but also to broader concerns related to the sustainability of journalism.

A. INCLUSIVE AND TRANSPARENT DATA CURATION

Teams make choices as they train datasets. The choices they make about provenance and curation of datasets determine the quality and outputs of AI systems. Considering the significant influence these outcomes exert on the information space and the potential for human rights violations, these decisions

[83] Díaz, A. and Hecht-Felella, L. (2021). *Double Standards in Social Media Content Moderation, Brennan Center for Justice*. Available at: www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf (Accessed: 7 February 2024).

[84] Nicholas, G. and Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis,* Center for Democracy and Technology. Available at: https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Accessed: 7 February 2024).

[85] *ibid*.

should not be left to engineers and AI companies and entities alone. Moreover, they should not be left to teams that lack internal diversity.

There is a pressing need to actively involve civil society, researchers and affected communities in a sustained and meaningful way. An inclusive approach is essential to ensure that dataset choices reflect broader societal values, rather than technical or corporate priorities or the inadvertent bias of homogenous groups of engineers. Such a process should not be used by AI companies and entities to evade their responsibility to respect ethical principles in data provenance and curation practices, but rather provide the opportunity for stakeholders in shaping training datasets.

In addition, training datasets should be open to external scrutiny[86] to verify compliance with existing laws and regulations, identify potential issues, and enable rights holders seeking recourse for suspected misuse of their data to gather evidence and pursue legal action. This means that AI companies and entities should provide tiered transparency about data they use. While external researchers, watchdogs, and other vetted entities should have direct access to training datasets for investigative purposes (see Chapter 4, section 4.2), the general public should be able to understand some essential characteristics of the training datasets in a format that is accessible and user-friendly.

RECOMMENDATIONS TO **STATES**

> **Mandate AI companies and entities to provide information about their training datasets for public scrutiny in an easily accessible and understandable manner, including a searchable database.**[87] This disclosure should encompass:

- ♦ Provenance and composition of the data, including how data was obtained, selected, and whether it is licensed. This also includes information about the use of any copyright-protected or otherwise legally protected data.
- ♦ A summary of the demographic characteristics of the people whose data is included in the dataset.
- ♦ Metadata about the topics, cultural contexts, and languages covered by the dataset, including information about data volume and format.
- ♦ Known limitations and potential biases inherent in the datasets.
- ♦ A description of the data processing steps, including how the data was cleaned and prepared for training, and anonymization and aggregation techniques used.

[86] According to the draft EU AI Act, relevant national authorities should have full access to the training, validation and testing datasets used for the development of the high-risk AI system (Article 63.7a). Additionally, in special cases, the source code of the high-risk AI system can also be requested (Article 63.7b).

[87] Jernite, Y; (2023). *Training Data Transparency in AI: Tools, Trends, and Policy Recommendations*, Hugging Face Community Blogvvailable at: https://huggingface.co/blog/yjernite/data-transparency#data-transparency-in-focus-what-is-needed (Accessed: 7 February 2024).

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Set up an inclusive and participatory process that at a minimum includes equitable, sustained and substantive participation by independent researchers, civil society and affected communities. This process should be used to determine the rules and criteria guiding dataset provenance and curation.** This particularly involves establishing criteria for the inclusion or exclusion of potentially harmful and misleading data in training datasets for foundation models. Furthermore, ensure that these processes, rules, and criteria are transparent to the public.

> **Respect ethical principles and laws in data provenance and curation practices, which include at a minimum:**

- **Compliance with data governance, data protection and intellectual property laws in provenance and curation choices.**
- **Ensure that datasets curated do not discriminate** on the basis of race, color, sex, sexual orientation, gender identity, language, religion, political or other opinion, national or social origin, property, birth or other status.

RECOMMENDATIONS TO **STATES, INTERNATIONAL AND MULTILATERAL STANDARDS BODIES**

> **Convene multi-stakeholder discussions to establish standards and guidelines regarding the types of data permissible for training AI models,** including sensitive topics such as children's data, and potentially sensitive categories like search history data and travel logs that encompass such things as visits to specialized health clinics, LGBTQ+ related venues, offices of human rights organizations or political parties, and more. These discussions should also address corresponding reinforcement learning and red-teaming policies.

B. ACCESS TO HIGH-QUALITY DATASETS

The availability of high-quality data stands as a strategic imperative that underpins responsible AI development, acting as a bedrock to inclusive innovation.

States should play a central role, championing the development of high-quality public sector datasets, publishing them under open data standards, and supporting the development of datasets in underrepresented regions, cultures, and languages.

Current data curation practices have also highlighted the need for a robust data governance framework, providing comprehensive guidance on data scraping, collection, and utilization. This framework should prioritize the public interest, and grapple with intricate issues linked to copyright and privacy. Additionally, public authorities should help data owners and data users streamline data-sharing procedures, making available standardized tools and mechanisms that guarantee fairness and equity in data exchanges.

RECOMMENDATIONS TO **STATES**

> **Cultivate the development and availability of high-quality public sector datasets and the implementation of open data standards.** This includes the production of data, the cleaning of datasets, and making them available under open data standards in a machine-readable and reusable format.[88]

> **Encourage the development of datasets in underrepresented countries, cultures and languages by allocating resources and support. This includes research grants, civil society funding, calls for contributions, data donations,[89] and support to media as a creator of data** (see subsection f). These efforts should focus on including a wide range of voices and representing historically marginalized communities, ensuring that AI systems align with and reflect the diverse cultures and values present within the country or community.

> **Develop a framework with clear rules on data governance. Such a framework should provide clear rules for data scraping, data collection and data use.** The use of data and access to datasets should be based on the public interest of the AI project, on the quality, legality, and sensitivity of the datasets, and on the amount of data involved. There should be a presumption of illegality for scraping data for inclusion in training datasets unless the data is published under open data standards, or AI developers and deployers can demonstrate the legality of their use of datasets. This also requires updated definitions of data sharing licenses to provide legal clarity, such as creative commons (which were elaborated prior to widespread AI data-scraping practices). In general, data minimization and data protection should govern technological companies' activities, meaning they should only collect the minimal data needed to run their products.

> **Develop data-sharing frameworks for organizations holding data and organizations seeking to use it to develop AI.**[90] For data to be shared, data owners and consumers often need to negotiate terms on a case-by-case basis. Developing standardized tools, templates, and practical mechanisms for data sharing would not only overcome practical obstacles, but also clarify how to assign ownership or rights related to the outputs generated from this shared data (e.g., AI models, analytical insights, research findings) and allocate benefits. Ultimately, these frameworks would guarantee that data exchanges are safe, equitable, and fair for all parties involved. Given the importance of trust in data-sharing practices, these frameworks should be developed and managed by an independent organization with a well-established reputation for data security and effectiveness. This organization should assume a pivotal role, exercising key functions such as providing the framework to define the data to be shared, mediating the data's intended purpose and use, establishing data transfer and storage protocols, and clarifying how any commercial value generated should be distributed.

[88] *The FAIR Guiding Principles for scientific data management and stewardship* can provide a framework for publishing such datasets. Available at: www.nature.com/articles/sdata201618 (Accessed: 7 February 2024).

[89] "Data donations" refer to a practice where individuals, public organizations, companies, or entities voluntarily contribute their data for research, development, and other uses in the public interest. In the context of AI development, data donations can help enrich datasets, especially for underrepresented countries, cultures, identities, and languages.

[90] This recommendation is based on the Data Trusts proposal advanced to the UK Government: Wendy Hall, D. and Pesenti, J. (2017). *Growing the Artificial Intelligence Industry in the UK*. Available at: https://assets.publishing.service.gov.uk/media/5a824465e5274a2e87dc2079/Growing_the_artificial_intelligence_industry_in_the_UK.pdf (Accessed: 7 February 2024).

- **Ease access to copyrighted material to be used as training data for public interest research, while also protecting intellectual property and user privacy.**
- **Encourage AI companies and entities to open up their training datasets for the use of public interest research and alternatives to for-profit AI systems.**

C. ADDRESSING BIAS

AI systems are frequently trained on large datasets that predominantly feature Anglocentric and Eurocentric perspectives. Furthermore, communities that are historically marginalized are often under- or misrepresented in these datasets – as well as in the teams who gather the data and train the systems. The diversity challenges in the data and on teams create bias that manifests in the AI systems.

Pervasive bias in training datasets and teams can result in lack of diversity in the perspectives and content AI systems understand and generate. In some cases, it can also lead to detrimental outcomes that disproportionately affect underrepresented cultures, as the systems may not adequately recognize or appropriately respond to diverse cultural contexts and nuances. For example, it has been found that the over-representation of Western viewpoints in datasets can result in AI systems perpetuating stereotypes and generating inappropriate or offensive content, particularly when it comes to representing women and non-Western cultures, with results often reflecting hyper-sexualized, misogynistic, and even White-supremacist ideologies.[91]

In light of these issues, addressing bias in training data and teams becomes an imperative to counteract the risk of dominant cultural hegemony.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

- **Ensure teams working on training datasets are inclusive and diverse.** This means championing a diverse workforce, training staff on inherent bias, strengthening diverse recruitment pipelines, and collaborating with education authorities to encourage a diversity of people to join the field.
- **Enrich training datasets to address issues of bias and representation of diverse cultures. This includes:**
  - **Diversifying the data used for training AI systems by collecting authentic data from underrepresented groups, regions, cultures and languages**. This involves directly sourcing data from groups, regions, cultures and languages that are currently underrepresented in datasets, recognizing the importance of fair compensation and ethical practices. One possible strategy to facilitate this collection is to organize “data donation campaigns” that empower communities to contribute to public interest projects while ensuring adequate compensation for their valuable data (see section b.).
  - **Weighting data segments differently.** Adjust the weights of different data segments appropriately to balance the dataset during training, giving higher weights to underrepresented or more critical segments. It is crucial that the

[91] Birhane, A. et al. (2021). *Multimodal datasets: misogyny, pornography, and malignant stereotypes*. Available at: https://arxiv.org/abs/2110.01963 (Accessed: 7 February 2024).

implications of this parameter adjustment must be understood to ensure that it will reduce bias without adversely affecting other performance measures.

- **Generating synthetic data.** This entails using artificial techniques, such as generative adversarial networks (GANs) and data augmentation techniques,[92] to create a more representative training dataset that approximates the true data distribution, particularly for the minority classes in imbalanced datasets. Synthetic data is particularly valuable in situations where existing data is biased or lacks diversity. While this method can mitigate privacy, copyright, and ethical concerns associated with using real data, it requires careful use. On one hand, synthetic data can help in addressing bias and enhancing privacy.[93] On the other hand, excessive reliance on synthetic data might affect the reliability of AI systems.[94]
- **Auditing and calibrating datasets.** This involves understanding the composition of the datasets and adjusting them to reflect a broader range of cultural and demographic backgrounds. This step helps in minimizing any biases that may be introduced into the algorithm.[95,96]
- **Proactively investing in developing technical solutions to construct datasets that accurately reflect the diversity of cultures globally,** with a particular focus on including underrepresented groups and smaller populations, to counteract biases inherent in standard benchmark datasets.

> **Conduct impact assessments to check for bias – including diversity and representation – before AI systems are deployed, and review them on an ongoing basis.** This could be done using internal "red teams" or third parties to audit data and models.[97]

RECOMMENDATIONS TO **STATES**

> **Mandate impact assessments by AI companies and entities to check for bias – including diversity and representation – before AI systems are deployed, and review them on an ongoing basis. This also involves auditing for bias when conducting independent audits** (see Chapter 4, section 2.3) **and enabling scrutiny by researchers** (see Chapter 4, section 4.2).

D. LOW-RESOURCE LANGUAGES

Recently, AI companies have been competing to incorporate the most languages into their multilingual language models. However, a key concern is how language models manage the "curse of multilinguality" – i.e., for a fixed model size, adding more multilingual data eventually starts to worsen the performance

[92] Shorten, C. and Khoshgoftaar, T.M. (2019). *A survey on Image Data Augmentation for Deep Learning.* Journal of Big Data, [online] 6(1). doi:https://doi.org/10.1186/s40537-019-0197-0. (Accessed: 14 Feb 2024).

[93] Zewe, A. (2022). *In machine learning, synthetic data can offer real performance improvements*, MIT News. Available at: https://news.mit.edu/2022/synthetic-data-ai-improvements-1103 (Accessed: 7 February 2024).

[94] Marwala, T. (2023). *Algorithm Bias — Synthetic Data Should Be Option of Last Resort When Training AI Systems,* United Nations University. Available at: https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems (Accessed: 7 February 2024).

[95] Cambridge Consultants (2019). *Use of AI in Online Content Moderation*. Available at: www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Accessed: 7 February 2024).

[96] European Parliament (2022). *Auditing the quality of datasets used in algorithmic decision-making systems*. Available at: www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf (Accessed: 7 February 2024).

[97] Silberg, J. and Manyika J. (2019). Notes from the AI frontier: *Tackling bias in AI (and humans),* McKinsey Global Institute. Available at: www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans (Accessed: 7 February 2024).

of the language model for both low-resource and high-resource languages.[98] Thus, companies developing large language models should focus not only on the number of languages included, but also on the quality of the model's performance in each individual language.[99]

Additionally, it is important to acknowledge that the development of natural language processing (NLP) capabilities unfolds in a cyclical manner. In the case of high-resource languages, such as English, this cycle tends to be virtuous: abundant, clean, human-annotated datasets lead to more advanced models and benchmarks, spurring further research, publications, and practical applications, which in turn drive demand for even more datasets. Conversely, low-resource languages are stuck in a vicious cycle due to the lack of tools, annotators, funding, and recognition in mainstream NLP publications and communities. To reverse this trend for low-resource languages, investment should focus on fostering self-sustaining scholarly NLP communities, requiring simultaneous efforts at all levels.[100]

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

- **Conduct impact assessments to identify inaccuracies and misrepresentations in the different languages in which the AI system will be available, before deployment, and review these assessments on an ongoing basis.** This could be done using internal "red teams" or third parties to audit data and models.
- **Establish clear metrics and benchmarks to evaluate the performance of language models in each language.**
- **Develop techniques to create more data-efficient language models that would require fewer resources to operate including:**
  - Developing models that prioritize specific domains or tasks.[101]
  - Using multilingual models for transfer learning from high-resource to low-resource languages.
  - Creating multilingual performance benchmarks.
- **Support the development of self-sustaining scholarly NLP communities by providing financial support and sharing low-resource language datasets used to train AI models, including (at least some) proprietary training data.[102]**

[98] Chang, T.A., Arnett, C., Tu, Z. and Bergen, B.K., (2023). *When is multilinguality a curse? Language modeling for 250 high-and low-resource languages,* arXiv. Available at: https://doi.org/10.48550/arXiv.2311.09205

[99] Nicholas, G. and Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis,* Center for Democracy and Technology. Available at: https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Accessed: 7 February 2024).

[100] *ibid*.

[101] Miller, K and Lohn, A. (2023). *Techniques to Make Large Language Models Smaller: AN Explainers*, Center for Security and Emerging Technologies. Available at: https://cset.georgetown.edu/publication/techniques-to-make-large-language-models-smaller-an-explainer/ (Accessed: 7 February 2024).

[102] Nicholas, G. and Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis,* Center for Democracy and Technology. Available at: https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Accessed: 7 February 2024).

RECOMMENDATIONS TO **STATES**

> **Invest in research into AI models for low-resource languages to address current market failures.** Government incentives can encourage NLP researchers to diversify their focus across various languages and approaches, rather than primarily concentrating on English, and to measure the impacts of AI models on different linguistic communities, linguistic preservation, and diversity efforts.[103]

> **Mandate internal impact assessments by AI companies to check for inaccuracies and misrepresentation in different languages before AI systems are deployed, and review them on an ongoing basis.** This also involves auditing for inaccuracies and misrepresentation when conducting independent audits (see Chapter 4, section 2.3) and enabling scrutiny by researchers (see Chapter 4, section 4.2).

E. PRIVACY AND DATA PROTECTION

Despite companies claiming to prioritize user privacy and safeguarding specific data points, this commitment generally does not extend to the insights derived from behavioral data, which provide the foundation for surveillance capitalism. In certain cases, companies deliberately employ dark patterns to prevent users from accessing robust privacy safeguards for their behavioral data. Such practices significantly hinder the effective exercise of user rights.

The limited commitment to privacy also extends to generative AI systems, which typically collect a vast amount of personal and conversational data, ranging from account and communication information to log and usage data, which might be shared with third parties. Users are unaware of the information they are sharing when they use AI systems to learn more about a financial situation they are in or health situation they are experiencing. This issue is particularly concerning for journalists who may need to protect sensitive information or sources.[104]

Conversations with AI interactive systems are often reviewed by AI trainers to improve the systems' performance and for research purposes. However, the rights of users regarding their data and the handling of their inputs and prompts are usually not clearly communicated. Furthermore, while users in some countries may have the option to request the deletion of their data from generative AI systems (the right to be forgotten[105]), they often lack the ability to delete specific inputs or prompts.[106] This lack of clarity and control leads to confusion and heightened privacy concerns.

[103] A good example of this is BigScience's BLOOM–an open-source LLM developed with public support (including funding from the French government)–able to generate text in 46 languages and 13 programming languages and perform tasks it hasn't been explicitly trained on. Both BLOOM and the model it was trained on (i.e., ROOTs, a 1.6TB multilingual dataset) are accessible for investigation by other NLP professionals. More info on BigScience's BLOOM are available at: https://bigscience.huggingface.co/blog/bloom (Accessed: 15 December 2023).

[104] Similar concerns might also apply to human rights activists and other stakeholders but are beyond the scope of the report and the mandate of the Forum on Information and Democracy.

[105] Enshrined in the Article 17 ("Right to erasure ('right to be forgotten')") in the European General Data Protection Regulation (GDPR).

[106] OpenAI (2023). Privacy policy. Available at: https://openai.com/policies/privacy-policy (Accessed: 15 December 2023)

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Develop and implement ethical business models and transparent systems that truly empower users to opt out of tracking and, to the greatest extent possible, exercise their right to be forgotten. This entails:**

- ♦ ceasing the collection of personal data during the AI user's interactions with the AI system;
- ♦ no longer disclosing the AI subject's personal data in outputs generated by the AI system; and
- ♦ whenever feasible, erasing any personal data concerning the AI subject from the training dataset utilized by the AI system.

> **Regarding users' inputs and interactions with an AI system, clearly communicate, in an easily accessible and visible way:**

- ♦ any data-processing activities for training and improving the model or other usages;

how the data is stored;

- ♦ whether the data is shared with third parties.

> **Explain how the use and storage of user inputs and interactions respects data protection laws and privacy.** This approach is essential to ensure that users can make informed decisions about their privacy and data usage.

> **Enable users to opt out of their input data and interactions being stored and used for AI system improvements.** This should be the default setting.

RECOMMENDATIONS TO **STATES**

> **Mandate that AI companies and entities demonstrate risk reduction in the use of protected data.**

> **Implement a system for lodging complaints about the use of protected data, ensuring the right to be forgotten, and putting the burden of proof on the company** (see Chapter 2, sections 1 and 4.3).

> **Mandate AI companies and entities to clearly communicate to users how their inputs and interaction data are used and stored, and to establish a mechanism allowing users to easily opt out of their data being stored and used.**

F. MEDIA SUSTAINABILITY AND COMPENSATION

The data scraped from the internet to train AI systems often also includes data produced and owned by media companies.[107] Currently, there is no transparency about the provenance of datasets, therefore the exact usage of media content in AI systems cannot be determined. In addition, it is unclear to what extent AI systems are fine-tuned and retrained on data inputted by users, including media organizations.

AI systems rely on high-quality data to train their systems. Often it is media organizations, and particularly public service media that own data in local and indigenous languages. There is thus a self-interest of AI companies and entities to ensure sustainability of media organizations to be able to access high-quality data from multiple cultural and linguistic groups. In addition, AI companies build their systems and thus their profit on the content provided by media organizations, raising the question of fair sharing of profit and of compensation.

There is a lack of legal clarity of copyright law and its application to AI, such as the American "fair use doctrine",[108] or the European Text and Data Mining (TDM) exception, to determine to what extent scraping media content would classify as a copyright infringement.[109] Media organizations are thus starting to opt out, indicating in their terms of service that their content cannot be scraped by AI companies without permission[110] and to sue AI companies for the unlawful use of their material.[111]

Governments have started adopting legislation to ensure big tech's contribution to media sustainability and compensation as is the case of the Australian Bargaining Code and the Canadian Online News Act. Principles for Fair Compensation were developed and adopted at the *Big Tech and Journalism – Building a Sustainable Future for the Global South* conference, in July 2023.[112] Similar approaches could be applied to AI companies as recognized in the Council of Europe's Guidelines on the Responsible Implementation of Artificial Intelligence Systems in Journalism.[113] Enforceable rights for compensation could be developed on different legal grounds such as copyright, competition law, or public interest taxation. They should respect the principles outlined in the Principles for Fair Compensation which include, among others, supporting public interest journalism, maintaining a plurality and diversity of media organizations, and putting in place a system that is based on fairness, sustainability and collective bargaining.

In an ideal scenario, public service media would contribute high-quality data to support building alternatives to for-profit AI systems.[114] Public interest media would be compensated for their contribution. Governments should therefore adopt legislation to enable media to be compensated for the use of their content. While voluntary deals between AI companies and media organizations can be a first step, they do not guarantee sustainability and stability for media organizations and might favor big media organizations.

[107] News/Media Alliance (2023). *White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement And Not a Fair Use*. Available at: www.newsmediaalliance.org/generative-ai-white-paper/ (Accessed: 15 December 2023).

[108] *ibid*.

[109] This also applies to other copyrighted content which goes beyond the scope of this report.

[110] See for example updated Terms of Service of the New York Times, available at: https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-Service (Accessed: 7 February 2023).

[111] Grynbaum, M.M. and Mac, R. (2023). *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*, The New York Times. Available at: www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (Accessed: 7 February 2024).

[112] Gordon Institute of Business and Science (2023). *Big Tech and Journalism - Principles for Fair Compensation*. Available at: www.gibs.co.za/news-events/news/pages/big-tech-and-journalism-principles.aspx (Accessed: 7 February 2024).

[113] Council of Europe (2023). *Guidelines on the responsible implementation of artificial intelligence systems in journalism (point 6.5)*. Available at: https://rm.coe.int/cdmsi-2023-014-guidelines-on-the-responsible-implementation-of-artific/1680adb4c6 (Accessed: 7 February 2024).

[114] Reporters without Borders Spinoza project is an example of such an approach. RSF (2023), *RSF and French general press alliance launch "Spinoza Project" to develop AI tool for journalists*. Available at: https://rsf.org/en/rsf-and-french-general-press-alliance-launch-spinoza-project-develop-ai-tool-journalists (Accessed: 9 February 2024).

RECOMMENDATIONS TO **STATES**

> **Establish enforceable rights for media organizations and journalists regarding the use of their content in AI training and application.** These rights should encompass:

- ♦ **A right to know,** allowing media organizations to demand transparency about the use of their content in training datasets and fine-tuning.
- ♦ **A right to opt out**, allowing media organizations to decide if their content can be used in AI systems. This includes the usage of content in training datasets, in grounding,[115] as well as in the use of data inputted by media organizations in AI systems when using the system.
- ♦ **A right to fair compensation whenever media content is utilized to train an AI system**. Possible options to guarantee the fulfillment of this right include:
  - Implementing a revenue-based tax on AI companies that specifically earmarks:
    - funds for redistribution to media organizations in national or regional contexts; and
    - a global tax to support media globally. The funds could be distributed through a global independent fund.
  - Targeted amendments to competition laws to address market power imbalances between AI companies and entities and media organizations, by allowing the latter to jointly negotiate the pricing, terms, and conditions under which their content can be used to train an AI system (i.e., a collective bargaining system for direct payment to media organizations).
  - Clearly defining that the use of media content in AI systems is a copyright infringement unless permission is given and compensation is paid.
  - Exploring the feasibility of implementing a royalty system similar to the compensation models used for artists on platforms like Spotify.[116] In this case compensation would apply not only during the AI training phase but also throughout the usage of the trained models.

> **Ensure that collected resources are allocated to media organizations through independent, transparent and accountable mechanisms that respect editorial independence and fairness, promote public interest journalism, plurality, diversity and sustainability of media organizations.**

> **In the long run, establish collaborative mechanisms between public service media and public alternatives to for-profit AI systems to fairly compensate public service media for data and content production while enabling the development of public alternatives to for-profit AI systems.**

[115] Grounding refers to using real-time search results as context for the AI systems responses, as defined by News/Media Alliance: News/Media Alliance (2023). *White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement And Not a Fair Use*. Available at: www.newsmediaalliance.org/generative-ai-white-paper/ (Accessed: 7 February 2024).

[116] Spotify (2023). *Royalties*. Available at: https://support.spotify.com/us/artists/article/royalties/ (Accessed: 7 February 2024).

1.2. HUMAN LABELING FOR AI TRAINING

Human data labeling plays a critical role in training various AI systems, including classifiers and generative models. This process involves humans rating the quality of the systems' responses and producing data, which is used as feedback for aligning AI's performance with both operational accuracy and societal norms. Alignment is particularly crucial for generative AI systems, ensuring they adhere to human values and ethical standards. This process is governed by comprehensive guidelines and best practices, which are essential to maintain the quality and integrity of the model training process.[117]

However, the specific guidelines and principles used in this process are not always transparent to the public. This lack of transparency also extends to the rationale behind specific decisions, the exact criteria used, and the values and motivations driving the human evaluators in the loop. Understanding these aspects is crucial for ensuring the trustworthiness and accountability of the AI system, as it directly influences how the model learns and evolves through human interaction.

Given the significant influence these decisions have on addressing bias, mis- and disinformation, hate, graphic violence, and other harmful content that can surface in AI outputs, AI companies and entities should promote openness and transparency in training methodologies (as well as the respect of labor laws). This transparency is essential to demonstrate their commitment to tackling these critical issues effectively. Moreover, AI companies and entities should actively involve a broad range of AI stakeholders in the process of defining these guidelines and best practices, and ensure that the AI's learning process respects and accurately reflects the diversity of human language and cultural expressions. Inclusive participation and a culturally sensitive approach are vital, as they ensure that a diverse array of perspectives and concerns are integrated in the refinement of AI systems, improving their alignment with the public interest.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Enhance transparency of human labeling processes. This entails making policies, guidelines, best practices, and procedures that guide human evaluators during the training process, transparent to the public.** This should include definitions used (such as about harmful speech), the areas/topics that are examined, and criteria used for accepting and rejecting specific outputs.

> **Implement participatory and inclusive processes, that at a minimum, include equitable, sustained and substantive participation by independent researchers and civil society for the formulation of policies, guidelines, best practices, and procedures guiding human labeling and alignment during the training process.** These should involve a broad range of stakeholders, including users, researchers, and relevant authorities. Importantly, civil society experts especially should be fairly compensated for their contribution and expertise (see Chapter 4, section 3.2).

> **Consider the contexts of diverse cultures and the nuances of language usage by different communities, particularly those that are historically marginalized, in the formulation of guidelines and best practices for human evaluators.** This consideration involves understanding and integrating into the human labeling framework cultural sensitivities, linguistic variations, and unique communication styles of different communities across the globe.

[117] OpenAI's (2023). *ChatGPT interaction on 5 December 2023.*

> **Implement a tiered system for human labeling processes, incorporating a participatory element from civil society.** This system would function as follows:

I. The content is initially reviewed by internal human evaluators at the AI company or entity. Their feedback is used to train and refine the AI model.

II. A carefully curated panel of community members – specifically including representation from historically marginalized and minority communities – are invited to review a sample of the content evaluated by the internal evaluators, with a focus on identifying harmful content impacting their specific community. Their feedback would then be used to further refine the AI model and improve its understanding of diverse contexts and sensitivities. Importantly, these community members should be fairly compensated for their contribution and expertise.

III. Implementation of a user-friendly mechanism within the AI system interface to facilitate flagging potential issues or concerns regarding the AI's outputs (see 2.3).

RECOMMENDATIONS TO **STATES**

> **Mandate AI companies and entities to enhance transparency of human labeling processes by making policies, guidelines, best practices, and procedures that guide human evaluators during the training process publicly available.**

1.3. CONTENT MODERATION AND RANKING SYSTEMS

Content moderation and ranking systems stand as gatekeepers of content, determining what information reaches users and what gets filtered out. They are used to assess the likelihood of content relevance, truthfulness, and adherence to platform standards, and to make decisions leading to downranking or removal.

For example, one study investigating the presence of racial bias in automatic hate speech detection models uncovered that these classifiers often over-moderated Black people, partly due to the classifiers' inability to understand nuances in context. Specifically, models were 1.5 times more likely to flag tweets written by self-identified Black users as offensive or hateful. Furthermore, tweets written in African-American English were "more than twice as likely" to be labeled as "offensive" or "abusive".[118] Training classifiers using biased data can exacerbate these issues, leading to skewed or biased AI outputs.

The design, training, and operation of content moderation classifiers and content ranking systems have fundamental implications for freedom of expression. Yet, the principles guiding their construction are often opaque, raising concerns about their fairness and the potential for systemic biases to be encoded into AI systems.

To effectively address these challenges, AI companies and entities should adopt a multifaceted approach grounded in transparency across all stages of AI development, and ensure the active involvement of

[118] Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N. and Allen, P. (2019). *The Risk of Racial Bias in Hate Speech Detection,* Association for Computational Linguistics, pp.1668–1678. Available at: https://aclanthology.org/P19-1163.pdf (Accessed: 7 February 2024).

civil society organizations (CSOs) and other relevant stakeholders in the refinement of systems used for content moderation and ranking.

At the same time, States or regulatory bodies should convene a multi-stakeholder process to define harmful content, ensuring consistent and efficient content moderation systems within their jurisdictions in line with international human rights law. In the long run, States could also consider building public training datasets for classifiers and ranking systems (see Chapter 3, section 1.5).

RECOMMENDATIONS TO **STATES**

- **Develop a shared definition of various kinds of harmful content among different companies and entities deploying classifiers.** This unified definition would serve as a guiding framework for the development and implementation of content moderation systems, ensuring consistency and coherence in addressing harmful content online at least within the same jurisdiction.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

- **Enhance transparency of content moderation classifiers and ranking systems. This entails:**[119]
  - ♦ Explaining what methods were used to create the datasets on which the systems are trained, and how the annotation process works.
  - ♦ Providing a detailed account of how content items are selected for training sets, and how annotators are selected to label these items.
  - ♦ Disclosing what role user reporting of harmful content plays in training set curation.
  - ♦ Regularly publishing comprehensive data regarding the performance of the AI systems, using standard measures that capture both false negatives and false positives.
- **Establish a testing regime** for content moderation and ranking systems. This regime should include diverse scenarios and inputs to evaluate the systems' performance across a wide range of content and cultural contexts, ensuring content is evaluated within the contexts in which it occurs.[120]
- **Provide avenues for public feedback and accountability** (see Chapter 4, section 3).

[119] GPAI (2023). *Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report,* Global Partnership on AI. Available at: https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Accessed: 5 December 2024).

[120] Cambridge Consultants (2019). *Use of AI in online content moderation, 2019 Report produced on behalf of Ofcom*. Available at: www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Accessed: 5 December 2024).

> **Adopt a community-driven approach, including language experts, civil society, local experts, heritage and language preservation advocates, linguists, human rights experts, and community members representing historically marginalized and minority communities, to annotator selection for content moderation, ranking and regression systems. This involves:**[121]

- Building training sets locally in collaboration with diverse community members. This process can be informed by the use of "citizens' juries".[122]
- Prioritizing annotators who are targets of the specific harms, to ensure accurate representation and understanding.
- Treating annotated datasets as a form of "case law" informing AI systems' decisions, allowing them to learn nuanced definitions of harmful content beyond textual definitions.

> **Establish a two-pass annotation process for classifying harmful content on social media platforms. In the first pass, annotators categorize content as "remove", "downrank", "untouched", or "uprank". This data is used to train a content moderation and ranking systems for discrete content categories. In the second pass, annotators rank pairs of content items based on harmfulness. This data is used to train a scorer that outputs a continuous harmfulness score for each item.**[123]

> **Train content moderation and ranking systems with "soft labels" to improve accuracy and confidence in decisions.** This involves using target distributions that reflect the variety of opinions among annotators rather than single, hard categories. By doing so, the AI system can learn from the variability in annotators' judgments and develop a better understanding of the nuances of content. Additionally, measuring the extent of disagreement among annotators on the categorization of a specific item can help an AI system determine how confident it can be regarding a moderation decision, thus ensuring more reliable outcomes.[124]

> **Incorporate disagreement metrics into downranking algorithms for more nuanced and effective moderation.** Downranking items likely to engender high disagreement less aggressively ensures a more balanced approach that acknowledges diverse perspectives and opinions.[125]

> **Open the market for harmful content classifiers to encourage competition, and simultaneously grant the possibility to choose among different classifiers.**

> **Whenever available, use external training datasets for a given category of harmful content created in a public domain, rather than creating a proprietary one behind closed doors.**[126] (see Chapter 3, section 1.5)

[121] GPAI (2023). *Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report,* Global Partnership on AI. Available at: https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Accessed: 5 December 2024).

[122] The concept of "citizens' juries" is borrowed from legal systems where a group of citizens collectively assesses the impact or harm of specific content, similar to making judgements in defamation trials. It is also a method used in participation processes, relying on a random selection system. As explained in: GPAI (2023). *Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report,* Global Partnership on AI. Available at: https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Accessed: 5 December 2024).

[123] *ibid*.

[124] *ibid*.

[125] *ibid*.

[126] *ibid*.

1.4. OPTIMIZATION OBJECTIVES

Understanding and carefully selecting optimization objectives in the development of AI systems is crucial to preserve the integrity of our digital information ecosystem. These objectives serve as a compass, guiding AI algorithms in determining what constitutes success in their operations.

Historically, AI systems, especially recommendation engines, have prioritized engagement, defined by some experts as "a set of user behaviors, generated in the normal course of interaction with the platform, which are thought to correlate with value to the user, the platform, or other stakeholders".[127] However, engagement is not always synonymous with actual personal or social value. For example, research suggests that more extreme or emotionally charged content tends to receive more interaction.[128] As a result, prioritizing engagement can lead to the promotion of harmful content such as misinformation, hate speech, and even ethnic violence, which can have far-reaching consequences on societal discourse and behavior, especially in high-stake situations such as in proximity to an election, conflict or a pandemic.[129] This is of particular concern in a digital space where the creation of bot accounts and synthetic content is becoming easier and easier, thus enabling these actors to manipulate the information space and to erode trust. In addition to alternative engagement metrics, this also requires, in the long-term, methods to enhance the visibility of authenticated content to strengthen trust in the online information space and reduce the risks of social bots polluting the space.

Although there have been calls for alternatives to engagement-based ranking systems, such as reverse chronological feeds, these alternatives come with their own limitations and do not necessarily encourage positive societal outcomes in the information space, nor trust in it. Therefore, platform operators should develop and experiment with alternative metrics for recommender and content moderation systems, which support an inclusive and trustworthy democratic information space.[130] This incentivizes the development and deployment of more ethical AI systems (see Chapter 3, section 1.5).

At the same time, researchers should investigate the impact of commonly used engagement metrics and possible alternatives on the dissemination of both illegal and legal-but-harmful content, including misinformation, conspiracy theories, hate speech, and harassment. Furthermore, researchers should investigate the impact of recommender systems on access to diverse and reliable information, on issues of public debate, and on high-stakes scenarios such as elections, conflict, war, and pandemics. Such investigation requires enabling external researchers to conduct experimental evaluations on social media platforms (see Chapter 4, section 4.3).[131]

Finally, the current set-up of AI algorithms and platforms encourages engagement on specific content pieces and their visibility, instead of fostering the constructive dialogue and exchange which are crucial for a democratic society. In the long run, AI companies and entities and platform companies should experiment with and develop alternative engagement mechanisms online.

[127] Bengani, P., Stray, J., & Thorburn, L. (2022). *Blog Post: What's Right and What's Wrong with Optimizing for Engagement*, Center for Human-Compatible AI at UC Berkeley. Available at: https://humancompatible.ai/news/2022/05/02/blog-post-whats-right-and-whats-wrong-with-optimizing-for-engagement/ (Accessed: 5 December 2023).

[128] *ibid*.

[129] Amnesty International (2022). *Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – Repor*t. Available at: www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ (Accessed: 5 December 2023).

[130] Bengani, P., Stray, J., & Thorburn, L. (2022). *Blog Post: What's Right and What's Wrong with Optimizing for Engagement*, Center for Human-Compatible AI at UC Berkeley. Available at: https://humancompatible.ai/news/2022/05/02/blog-post-whats-right-and-whats-wrong-with-optimizing-for-engagement/ (Accessed: 5 December 2023).

[131] Bengani, P., Stray, J., & Thorburn, L. (2022). *Blog Post: How to Measure the Effects of Recommenders. Understanding Recommenders,* Center for Human-Compatible AI at UC Berkeley. Available at: https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57 (Accessed: 5 December 2023).

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

- **Address the potential adverse effects of optimizing for engagement on societal discourse, behavior and democracy, especially in high-stakes situations.** This requires implementing a series of critical steps:
  - **Conducting a comprehensive threat assessment**. This should aim to:
    - **Analyze the impact of engagement metrics** (likes, shares, comments, click-through rates, and time spent on content), including through A/B testing, **on the amplification of harmful content related to social issues, elections, conflict and politics.**
    - **Identify which engagement metrics are most problematic** (i.e., likely to promote harmful content such as misinformation and disinformation, hate speech, or content inciting violence by analyzing data trends).
  - **Developing a contingency plan to adjust problematic engagement metrics in scenarios where they could lead to human rights violations, violence, or physical harm. This is particularly crucial in sensitive situations like elections, conflicts, and health emergencies.** This plan could include:
    - Alternative metrics to use in place of, or as substitutes for, metrics that have been identified as problematic.
    - Strategies to reduce the weight of problematic metrics in ranking algorithms.
  - **Engaging with a diverse group of stakeholders, including researchers, CSO representatives, disinformation experts, and relevant authorities, to solicit feedback and incorporate their perspectives into the threat assessment and contingency plan.**
  - **Testing the changes proposed in the contingency plan in a controlled environment.** This could involve A/B testing or deploying changes in specific demographic or geographies first.
  - **Establishing a system for continuous monitoring and refinement, ensuring transparency and easy access for users and trusted external parties to submit feedback.**
  - **Regularly reporting on the findings and adjustments made.** This step is crucial for ensuring accountability and enhancing the understanding among AI stakeholders of how various optimization metrics influence issues like misinformation, online hate, harassment, and violence against marginalized communities and possible strategies to minimize these threats.
- **Provide users with an easy and user-friendly opportunity to choose alternative recommender systems that do not optimize for engagement, but rank in support of positive individual and societal outcomes such as reliable information, bridging content or diversity of information.**
- **Initiate experimentation to test the effects of alternative engagement metrics,** which could include:
  - **User rankings** that rank content according to criteria such as reliable content, recommendability, clear and pedagogical, important and practicable, suitable for a

wide audience, resilient against backfiring, inclusive and diverse, encourages good habits.[132]

- **Consensual or bridging content**, meaning content with approval from diverse actors.[133] This requires an inclusive and transparent process to develop a system for providing approval to content.
- **Prioritizing "long-term engagement" (i.e., engagement duration) over "short-term engagement" (i.e., engagement intensity) as a metric in recommender systems.** Recent research suggests that optimizing for how intensively a user engages with content over a single session rather than how much time a user spends on the platform over an extended time period may lead to the amplification of polarizing content. However, when the emphasis is shifted to maintaining user engagement over time, recommender systems are more likely to support a balanced and less divisive content environment.[134]
- **Shifting focus from short-term engagement cycles, which favor recent content, to longer engagement cycles that also take older content into account, except for content and breaking news that are authenticated as reliable (such as by the Journalism Trust Initiative).** When it comes to recommending content, platforms differ widely in their bias toward recent content and events. Recent research suggests that platforms with extended engagement lifecycles benefit from having more time to moderate content before it reaches peak engagement. Conversely, platforms that are biased toward breaking information and events frequently face challenges in moderating content effectively within such short timeframes. As a result, platforms with longer engagement periods generally recommend less polarizing content than those operating on shorter engagement cycles.[135]

> **Apply design strategies proven to diminish the spread of harmful content.** In particular, AI developers and deployers should:

- **Reduce the use of engagement optimization for sensitive content including politics and health.[136]** This is known to reduce the spread of misinformation and polarizing content.
- **Increase the weight of metrics assessing the credibility and quality of news publishers and their reporting in ranking algorithms.** This is known to help ensure that authoritative and substantiated news stories are displayed more prominently, especially in proximity to an election.[137] The Journalism Trust Initiative offers a potential solution by implementing an international mechanism that rewards ethical journalistic practices. As this initiative was designed as an ISO standard, it holds the potential to be included in algorithms as a means to reward and prioritize good-quality journalism.[138] Collaboration with electoral management bodies can strengthen access to reliable election content during election periods.[139]

[132] Tournesol Project (Accessed: 7 February 2024).

[133] Bridging information is "content that meets approval (or generates positive engagement) across diverse groups of people." As defined in: Stray, J., Iyer, R., and Puig Larrauri, H. (2023). *The Algorithmic Management of Polarization and Violence on Social Media*, Knight First Amendment Institute at Columbia University. Available at: https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media (Accessed: 15 December 2023).

[134] Edelson, L., Haugen, F., and McCoy, D. (2023). *Demystifying Social Media Feeds: A Taxonomy and Transparency for Algorithmic Feed Systems Designs.* Draft manuscript.

[135] *ibid*

[136] Stepanov, A. and Gupta, A. (2021). *Reducing Political Content in News Feed*, Meta. Available at: https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/ (Accessed: 8 February 2024).

[137] Lyons, K. (2020). *Facebook rolls back 'nicer' News Feed that boosted mainstream publishers*, The Verge. Available at: www.theverge.com/2020/12/17/22180259/facebook-news-feed-change-post-election-publishers-misinformation (Accessed: 8 February 2024).

[138] Journalism Trust Initiative available at www.journalismtrustinitiative.org/

[139] Forum on Information & Democracy (2024). *Protecting Democratic Elections through Safeguarding Information Integrity*, Forum on Information & Democracy, International IDEA, Democracy Reporting International. Available at https://informationdemocracy.org/2024/01/30/tech-firms-governments-urged-to-combat-digital-election-threats/ (Accessed: 8 February 2024)

- **Initiate experimentation to change the underlying structure of online engagement to encourage constructive debate and trust**. This can be achieved by:
    - Enhancing trust in online content by implementing an opt-in mechanism for author authentication, utilizing privacy-preserving cryptographic techniques such as zero-knowledge proofs.[140] Such a system should not be used to strengthen censorship.
    - Amplifying authenticated content from authors who choose to be verified.
    - Providing users with a provenance trail that details its origins and the authentication status of its authors, enhancing transparency and credibility.
    - Implementing alternative ranking systems that attempt to create constructive dialogue, compromises and exchange.

RECOMMENDATIONS TO **STATES**

- **Implement a mechanism to ensure that vetted independent researchers and external auditors can run on-platforms experiments[141] such as A/B tests to unveil causal relationships between algorithm designs and parameters and negative outcomes such as the spread of harmful and polarizing content** (see Chapter 4, section 4.3).[142]
- **Hold platforms accountable for reducing the weight of optimization metrics that are known to lead to negative individual and societal outcomes.**
- **Mandate AI companies and entities to provide users with the opportunity to choose alternative recommender systems that do not optimize for engagement but rank in support of positive individual and societal outcomes.**

1.5. CONTENT AUTHENTICITY AND PROVENANCE

The rapid development of AI content generation and fast distribution of such content makes differentiating AI-generated versus authentic content more challenging. Hence, discussions around solutions are pivoting towards the development of enhanced mechanisms for ensuring content authenticity and provenance.

Content authenticity and provenance is the assurance that digital content accurately represents its origin and has not been (maliciously) manipulated. It involves a traceable life cycle of the content, having a record of its creation, modification, and distribution. There have been initiatives in developing and documenting standards for content authenticity and provenance mechanisms of digital media. For instance, the Content Authenticity Initiative (CAI) is an organization that has developed a tool using cryptographic hashing to provide verifiable, tamper-evident signatures on digital contents,[143] and allowing consumers to view historical information about the content. A further organization, Project Origin, was established to provide a platform to discuss creation and adoption of a new media

[140] Zero-Knowledge Proof are methods to verify claims without disclosing the actual information as defined in Aad, I. (2023). Zero-Knowledge Proof. In: Mulder, V., Mermoud, A., Lenders, V., Tellenbach, B. (eds) *Trends in Data Protection and Encryption Technologies* . Springer, Cham. available at: https://doi.org/10.1007/978-3-031-33386-6_6 (Accessed: 8 February 2024)

[141] Such experiments must respect data privacy and include notification of users as discussed in Chapter 4, section 4.3.

[142] Bengani, P., Stray, J., & Thorburn, L. (2022). *Blog Post: How to Measure the Effects of Recommenders. Understanding Recommenders,* Center for Human-Compatible AI at UC Berkeley. Available at: https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57 (Accessed: 5 December 2023).

[143] Content Authenticity Initiative (n.d.). *How it works*. Available at: https://contentauthenticity.org/how-it-works (Accessed: 7 February 2024).

provenance tracking process, initially for news and information content, among a set of partners.[144] A third organization, the Coalition for Content Provenance and Authenticity (C2PA) brings together the efforts of the CAI and Project Origin[145]. Watermarking, or embedding signals to AI generated contents, is another tool that is extensively employed to establish content authenticity.

As the technologies become more advanced, so attacks are becoming more sophisticated. For instance, the emergence of tools capable of removing or circumventing watermarks underscores the need to develop more robust and tamper-resistant watermarking methods. The table below is a simulation of potential type of attacks, method of attacks, and combinations of possible defenses, which can further be developed to more effective tools and solutions. Aside from content authenticity and provenance mechanisms, defenses include improving user education to help discern AI-generated contents, regular model evaluation to assess output reliability, detection algorithms, more stringent criteria for content amplification, and other adversarial training within the systems.

Table 1.1: Overview of potential attacks on the information space, the methods employed, and possible defenses

ATTACK	METHODS	DEFENSES
Actor unintentionally creates and/or shares misleading AI-generated content	Fabricated news/information AI hallucinations AI-generated images	• User education • Content provenance • Detection mechanisms
Actor intentionally creates misleading/fake content but does not have intent/capabilities to circumvent watermarking systems/detectors	Humor, parody etc. Non-consensual intimate images Spam or "content farms" where having some of the content detected as fake does not matter to the actor (whether for financial or political motivations).	• Watermarking • Content provenance • Detection mechanisms
Actor intentionally creates misleading/fake content and circumvents detectors	Removing or circumventing watermarks.	• Tamper-resistant watermarks • Detection mechanisms
	Circumventing detector through methods which are not based on watermarks (e.g., adding noise to break pixels or adding random irrelevant information content, etc.) to make it more challenging to detect manipulations.	• Content provenance
Actor intentionally creates misleading/fake content and forges provenance	Modifying real content and forging provenance or authenticity data. (Adding fake logos or labels, cryptographic attacks against signatures, stealing signing keys, adding fake logos or labels.) Making fake content and forging provenance. (Showing a fake image/scene to a real camera, adding fake logos or labels, cryptographic attacks, stealing signing keys.)	• User education • Watermarking • Content provenance • Content amplification / non-recommendability system
Actor attempts to produce model failures	Producing specific generative model outputs (e.g., poisoning public datasets, contributing bugs to open-source generators, or compromising private datasets). Causing detection model failures (e.g., contributing bugs to open-source watermark creation or detection).	• Robust model design • Adversarial training • Regular model Evaluation/audit

[144] Project Origin (n.d.). Project Origin. Available at: www.originproject.info/about (Accessed: 7 February 2024).

[145] At the time of writing, Meta, Google and OpenAI just announced joining C2PA and implementing its standards, alongside Microsoft who is among its founders.

Actor attempts to alter platform behavior (via hacking etc.)	Platform manipulation without internal access, e.g., brigading, bots, etc. Preventing platform detectors from running, or replacing detectors with broken detectors. Adding misleading labels or other user interface changes. Alterations to content ranking or content moderation.	• Watermarking • Content provenance • Detection algorithms • Content amplification / non-recommendability system

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Develop robust and reliable provenance-indicating watermark and detection techniques (such as watermarks or content logging) for content generated by their AI systems, and make these techniques publicly and freely available.**[146]

> **Collaborate with researchers to develop and implement effective watermarking and detection tools.**

> **Implement recognised content authenticity and provenance tools and standards. For content creation and editing systems, this includes generating digital signatures and other provenance metadata.**

RECOMMENDATIONS TO **STATES**

> **Establish a comprehensive support program for research focused on improving methods for detecting AI-generated content such as watermarking or content logging. This could include access to advanced technical infrastructure, funding, and collaboration with platforms.** In terms of research efforts, these should build upon existing literature and further research on open questions and current limitations for generative (i.e., exploring diffusion models[147]) and discriminative tasks, language models,[148] and fake content generated by generative adversarial networks (GANs).[149]

> **Establish comprehensive standards on invisible provenance-indicating watermarking and visible labeling requirements through a participatory process. This includes:**

♦ Invisible provenance-indicating watermarking requirements to be included in each AI system deployed.

♦ Visible labeling of AI-generated content (see Chapter 2, sections 3.2 and 4.1).

♦ Details about the technical requirements.

[146] The draft EU AI Act mandates providers of AI systems including GPAI systems that generate synthetic audio, image, video or text content to ensure the outputs of the AI system are marked in a machine readable format and detectable as artificially generated or manipulated (Article 52).

[147] Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M. and Lin, M. (n.d.). *A Recipe for Watermarking Diffusion Models*, arXiv. Available at: https://arxiv.org/pdf/2303.10137.pdf (Accessed: 7 February 2024).

[148] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. and Goldstein, T. (n.d.), *A Watermark for Large Language Models*, arXiv. Available at: https://arxiv.org/pdf/2301.10226.pdf (Accessed: 7 February 2024).

[149] Yu, N. et al (2022). *Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data*. Available at: https://arxiv.org/abs/2007.08457 (Accessed: 7 February 2024).

- **Establish standards governing content authenticity and provenance, including on author authentication, through a participatory mechanism, including civil society and academia. This could start by building upon the processes, standards and technical solutions developed by the Coalition for Content Provenance and Authenticity. The aim would be to establish a recognized standard and to create consistency among AI systems on content provenance and author authenticity.**
- **Mandate platforms to detect provenance information and AI-generated content** by the best means currently available and **to display this information to the end user.**

RECOMMENDATIONS
TO **PLATFORMS**

- **Implement existing standards on content provenance (such as the ones developed by the Coalition for Content Provenance and Authenticity), detection and watermarking to strengthen information integrity online and enable greater visibility for users. For platforms, browsers, and other end user applications which display content, this includes checking for provenance metadata and displaying it to the user, to help them recognize authenticated content.**

2. TESTING AND MITIGATING RISKS OF AI SYSTEMS

As previously noted, AI systems can inadvertently perpetuate harmful biases, breach legal frameworks on privacy and data protection, and even threaten democratic values and human rights, including by facilitating disinformation and eroding privacy. These risks were recently brought into the spotlight by the G7, which called on AI companies and entities to take appropriate measures throughout the AI lifecycle – including development, deployment, and market placement – to identify, evaluate, and mitigate these risks.[150]

To properly safety-test new AI systems to be deployed in the information space and implement the commitments contained in the voluntary International Code of Conduct[151] agreed by G7 leaders, AI companies should expand red-teaming tests to cover a broader range of languages and cultural frameworks, thus reflecting AI's global usage. Moreover, they should actively engage with external researchers and civil society organizations, who can help identify biases and ethical concerns that

[150] European Commission (2023). *Hiroshima Process International Code of Conduct for Advanced AI Systems*. Available at: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Accessed: 8 February 2024).
[151] *ibid*.

AI developers may overlook. To ensure that this crucial practice goes beyond a checkbox exercise, engagement should be carried out in a clear, sustained, substantive, and compensated manner.

Pre-market risk assessments are also critical to uphold democratic values and human rights. To be effective, risk assessments should focus on specific use cases, and rigorously evaluate risks related to content moderation, exposure to illegal or harmful content, and the spread and impact of misinformation and disinformation, particularly concerning electoral integrity. At the same time, they can proactively identify and mitigate risks from the outset, and continuously monitor and address potential shortcomings throughout the whole AI lifecycle. Publishing "model cards" that detail AI uses, mitigation strategies, and descriptions of training and testing processes can enhance ongoing risk monitoring and mitigation and promote accountability.

Lastly, even with rigorous pre-launch assessments and conformity checks, AI systems require ongoing vigilance to remedy potential data and model poisoning, to address unforeseen real-world outcomes, ensure compliance with evolving regulations and standards, and to swiftly implement remedies when deviations occur. In this context, post-market monitoring can effectively protect AI systems from data- and model-tampering and ensure their ongoing compliance with changing regulations. Implementing a structured user feedback mechanism, protocols for rapid issue resolution, collaboration with trusted flaggers, and regular AI system updates is also essential.

2.1. RED-TEAMING

Although definitions and goals of AI red-teaming activities vary substantially, there is a broad consensus that AI red-teaming activities are a key component of AI risk management.[152] Notably, documents such as the White House Executive Order,[153] the G7 Principles,[154] and the UK Government's *Emerging Processes for Frontier AI Safety*,[155] released ahead of Bletchley Park Declaration, have all recognized red-teaming as a preferred method for managing AI risk.

As the AI community works towards a unified definition of red-teaming, it is also crucial to devise targeted strategies in current red-teaming efforts to identify potential vulnerabilities in generative AI systems intended for use in the information space.

First, red-teaming tests should be broadened beyond English-speaking and Western contexts to lower the risk of bias and misuse. Second, the composition of red teams should mirror the diversity of users and comprise a wide range of subject matter expertise. Third, red team members should collaborate with independent external researchers. Fourth, AI developers and deployers should try to make AI more explainable by publishing AI model cards before starting to develop/deploy AI systems, and potentially have those evaluated before taking action. Fifth, it is essential to implement democratic or participatory processes to define the guidelines for red-teaming, be transparent about these rules, and create mechanisms for continuous, external red-teaming of AI models for risks to rights and safety.

While immediate implementation of these actions by companies should not be delayed, there is also a pressing need for international standards bodies to establish mandatory standardized red-teaming

[152] Frontier Model Forum (2023). *Frontier Model Forum: What is Red Teaming?* Available at: www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf (Accessed: 7 February 2024).

[153] The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Available at: www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Accessed: 7 February 2024).

[154] European Commission (2023). *Hiroshima Process International Guiding Principles for Advanced AI system*. Available at: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system (Accessed: 7 February 2024).

[155] UK Government (2023). *Emerging Processes for Frontier AI safety*. Available at: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (Accessed: 8 February 2024).

practices for AI systems. The range of activities in current red-teaming practices for generative AI systems includes techniques such as hacking to extract sensitive data, prompt injection, and system manipulation to generate harmful outputs such as cyber malware, virus toxic chemicals formulas, and terror attack strategies. These activities present significant risks, notably their potential for adversarial exploitation. Without addressing this regulatory gap, unregulated red-teaming efforts could represent a severe existential threat to public safety and the integrity of democratic systems.

RECOMMENDATIONS TO AI COMPANIES AND ENTITIES

> **Broaden the scope of red-teaming tests to cover a wide range of languages and cultural frameworks.** This expansion will enhance red teams' capacity to uncover vulnerabilities in AI systems that go beyond the predominant English and Western contexts.

> **Develop red-teaming guidelines in a participatory and inclusive manner that** at a minimum includes equitable, sustained and substantive participation by independent researchers, civil society and affected communities and **make these guidelines publicly available.**

> **Provide relevant authorities with detailed information regarding the performance results of AI models in red-teaming testing. This information should also be accompanied by a description of the measures implemented to address identified issues and enhance the overall security of the system.**[156] The detail of the information required is dependent on the systemic risks the AI system poses to the information space (see 2.2)

> **Adequately staff red teams to ensure that their composition mirrors the diversity of users in terms of demographics, languages, and cultural backgrounds. Furthermore, include representatives from marginalized and vulnerable communities who are often most susceptible to harm.**

> **Ensure that red team members possess expertise across a wide range of domains, including:**

- ♦ Human and civil rights
- ♦ Journalism ethics and standards
- ♦ Election integrity
- ♦ Data protection
- ♦ Intellectual property
- ♦ Mis- and disinformation, online hate, harassment, extremism, and illegal speech

> **Promote collaboration with external researchers in red-teaming activities to bridge knowledge gaps and prevent groupthink, recognizing the importance of fair compensation and ethical practices.**

> **Conduct red-teaming testing both at the model level and at the application level** to ensure comprehensive risk assessment and mitigation across all stages of AI system development and deployment.

[156] The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Available at: www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Accessed: 7 February 2024).

> **Invest in establishing norms and mechanisms for ongoing, external red-teaming of foundation models and their applications.** External red-teaming has proven to be an effective method for uncovering not only emerging AI safety and security risks, but also key issues such as bias, discrimination, and privacy.[157]

RECOMMENDATIONS
TO **STATES**

> **Support collaborative efforts to establish comprehensive guidelines for conducting red-team tests of AI models.** These guidelines should encompass procedures and methodologies to ensure rigorous testing and evaluation, and should address issues such as bias, discrimination, mis- and disinformation, hate speech, and other content that might infringe upon human rights.

> **Mandate AI companies and entities to provide relevant authorities with detailed information regarding the performance results of AI models in red-teaming testing, and hold them accountable for conducting red-teaming in line with established guidelines.**

> **Mandate AI companies and entities to make their red-teaming guidelines publicly available.**

2.2. PRE-RELEASE RISK ASSESSMENTS

Although red-teaming can help identify AI system vulnerabilities, it also has limitations when it comes to assessing fluid categories of harm in the information ecosystem, such as misinformation and hate speech.[158]

Additional measures, such as comprehensive pre-release AI risk assessments[159], can serve as a crucial guardrail in this space, ensuring ethical development and comprehensive assessment of AI systems impacting the information space before their release. These assessments also serve to determine if a system is high, medium, or low risk and, consequently, the regulations with which the AI systems need to comply (see Chapter 4, section 2).

Systemic risks to the information space stemming from AI systems include:

- Risks associated with the dissemination of illegal content.
- Risks impacting the exercise of fundamental rights, including freedom of expression and of information, including the right to access information; media freedom and pluralism; the right to vote and political participation; the right to private life; data protection, and the right to non-discrimination.

[157] Mislove, A. (2023). *OSTP Blog Post: Red-Teaming Large Language Models to Identify Novel AI Risks,* The White House, Office of Science and Technology Policy. Available at: www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/ (Accessed: 7 February 2024).

[158] Robertson, D. (2023). *When 'red-teaming' AI isn't enough*, POLITICO. Available at: /www.politico.com/newsletters/digital-future-daily/2023/10/25/when-red-teaming-ai-isnt-enough-00123577 (Accessed: 7 February 2024).

[159] To compare, the draft EU AI Act mandates high-risk AI systems to conduct an assessment of the impact on fundamental rights before putting systems on the market (see Article 29a).

- Risks impacting democratic processes, civic discourse and electoral processes, and other high-stakes scenarios such as public health and peace.[160]

Table 1.2 provides an overview of the potential harms that can arise from the use of AI systems within the information and communication space, and that can pose a systematic risk. It also provides ideas of factors to be taken into account to assess the likelihood of such harm occurring and its severity. The table needs to be regularly updated in line with the latest technological developments and associated arising risks.

Table 1.2: Assessing systemic risks of AI systems to the information space

POTENTIAL HARMS TO BE ASSESSED IN THE RISK ASSESSMENT	FACTORS TO BE TAKEN INTO ACCOUNT TO ASSESS THE LIKELIHOOD AND SEVERITY OF THE HARMS OCCURRING[161]
• Copyright violations • Privacy and data protection violations • Authenticity and trustworthiness challenges • Creation and dissemination of misinformation and hallucinations • Creation and dissemination of disinformation and deepfakes • Creation and dissemination of illegal speech, hate speech, and violence • Bias, discrimination, and cultural hegemony in AI-generated content • Surveillance and data exploitation • Manipulation, deception, and impersonation • Hyper-personalization • Polarization and conflict escalation • Censorship • Harassment	• Intended purpose of the AI system • Capability of the AI system to create harm (results of red-teaming, quality of datasets, etc.) • Capability of the AI system to react to harms and correct them (risk-mitigation measures, feedback and complaint mechanisms, correction procedures) • Transparency of the AI system and external scrutiny • Safety and robustness of the AI system against abuse (i.e., cybersecurity) • Accessibility of the AI system to the general public • Accessibility of the AI system to bad actors • Number of actual and potential users • Use of the AI systems by critical actors of the information space (i.e., media, government) • The type and amount of the training data • Capability of AI systems to act autonomously • Past record of the harm caused by the AI system

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Assess the systemic risks to the information space stemming from the utilization of AI systems in content verification and moderation, content curation and recommendation, ad targeting and delivery, content creation, and content personalization.** Seek input from relevant stakeholders such as CSOs, researchers and affected communities while compensating them fairly for their expertise.

- ♦ **Include not only theoretical assessment of risks, but also focus on specific use cases** in risk assessment procedures and reports**.**

> **Develop "model cards" for high-risk AI systems and make them available to regulators. This should apply exclusively to new AI systems and substantive changes to existing ones. The model cards should be published before starting to develop the model, evaluated before release and regularly updated to enhance transparency. These model cards should detail:**

- ♦ **potential misuses of the AI system;**

[160] Inspired by the DSA systemic risks, Article 34.
[161] See criteria to assess risks of harm from AI systems in the draft EU AI Act, Article 7.

- **mitigation strategies;**
- **planned training and testing processes – including the training data – prior to and during system development.**

RECOMMENDATIONS
TO **STATES**

> **Mandate AI companies and entities to publish "model cards" for high-risk AI systems before starting to develop an AI model, and to periodically update them.**

> **Mandate AI companies and entities to conduct systemic risk assessments prior to release of AI systems, and to make the results of these risk assessments transparent, auditable, and publicly accessible** through the public repository (see Chapter 4, sections 2.1 and 4.1). This approach ensures a clear and accountable record of the potential risks associated with each AI system.

2.3. POST-RELEASE MONITORING SYSTEMS AND RISK-MITIGATION MEASURES

Even with thorough pre-release assessments in place, AI systems require constant vigilance. Pre-release testing, though crucial, cannot entirely predict the complexities and unexpected scenarios encountered in real-world applications.

A primary concern is the continuous threat of data and model poisoning. Malicious entities could manipulate data or alter AI models, undermining their integrity and leading to biased or incorrect outcomes. Additionally, the scope of pre-release risk assessments might not encompass the entire range of real-world behaviors and decisions of AI systems, necessitating continuous monitoring to identify and mitigate emerging risks.

These potential threats and limitations can result in unforeseen risks and unintended consequences, underscoring the importance of robust post-release monitoring systems, including routine updates for systemic risk assessments and periodic maintenance. In addition, effective post-release monitoring must be supported by adequate risk mitigation and output moderation measures. These include the establishment of structured mechanisms for users and trusted flaggers to report issues and concerns, partnerships with fact-checkers, as well as protocols for swift response and rectification in case of identified risks or failures.

While similar mechanisms are already required for Very Large Online Platforms (VLOPs) in Europe under the Digital Services Act (DSA) and the UK's Online Safety Act, a similar risk mitigation and output moderation mechanism should also apply to generative AI systems, ensuring that undesirable or harmful outputs and content are identified and treated and rights respected and protected.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

- **Implement risk mitigation and output moderation measures to continuously update the system and respond to identified risks. They should include:**
  - **Notice and action mechanisms for capturing user feedback for private** (see Chapter 2, sections 2 and 3.1) **and publicly available feedback** (Chapter 4, section 3.1).
  - **Complaint mechanisms.**
  - **Collaboration with trusted flaggers and fact-checkers.**
  - **Continuous red-teaming to tackle identified risks.**
- **Update the systemic risks assessments regularly post release of the AI system and modify the AI system according to newly identified risks. Make the results of these assessments publicly available.**

RECOMMENDATIONS TO **STATES**

- **Mandate deployers of AI systems to put in place risk-mitigation measures which include:**
  - **Notice and action mechanisms for capturing user feedback** (see Chapter 2, sections 2 and 3.1).
  - **A complaint mechanism for appealing against generated output that infringes upon existing rights, including clear pathways for redress (**see Chapter 2, sections 3.1, 3.2 and 4.3).
  - **Protocols for rapid response and resolution in case of identified issues.**
  - **Continuous red-teaming.**
  - **A structured collaboration with trusted flaggers and fact-checkers.**
- **Mandate deployers of AI systems to conduct regular updates of systemic risk assessments, along with system maintenance and updates in accordance with the assessment results. The frequency of these updates should depend on the number of users of the AI system and its systemic risk to the information space. These assessments should take into consideration the problems that have been identified and flagged by users** (see Chapter 4, section 3.1)**, and they should be made publicly available.**

CHAPTER 2:
LIABILITY AND ACCOUNTABILITY REGIMES

INTRODUCTION

AI systems can potentially yield erroneous, inaccurate, biased, or discriminatory outcomes. They have the very real potential to infringe upon existing laws such as privacy, data protection, intellectual property, and anti-defamation or hate speech provisions. And importantly, they remain vulnerable to misuse – from manipulating public opinion to spreading mis- and disinformation or generating deepfakes and other doctored content.

As AI systems increasingly integrate into and shape our information and communication space, fundamental challenges arise. These challenges impact liability and accountability regimes that define the social contract between developers and deployers of AI systems and democratic society.

First, the presence of a wide array of actors in the AI ecosystem creates a 'problem of many hands', where responsibility is diffused among many participants and across the value chain. Second, the interaction between humans and AI systems, often referred to as 'humans in the loop', adds another layer of complexity to assigning responsibility.[162] *A key issue is to what extent decision-making can be confined to AI systems. Finally, the interaction between multiple AI agents, particularly when it occurs at high speed and on a vast scale, can result in unpredictable and potentially harmful outcomes. Even after extensive testing and fine-tuning, the interplay of individual AI agents, each capable of learning and iterative improvement, presents risks that are difficult to predict or control.*[163]

These scenarios all challenge the attribution of accountability and liability of AI decisions and outputs. Addressing these issues is crucial to provide AI users and subjects with adequate legal protection against human rights and potential criminal violations, thereby preserving trust in AI technologies used to create, disseminate, and consume information. The Council of Europe's Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law[164] *equally recognizes the need to establish accountability for human rights violations and effective remedies.*

This chapter explores how policymakers should assign responsibility along the whole AI value chain and ensure that AI developers and those deploying and using AI systems take these responsibilities seriously. It then highlights the need to reverse the burden of proof in liability cases involving AI systems. Additionally, it outlines measures to empower AI subjects to hold responsible parties accountable for the harms caused and damage inflicted.

[162] Yeung, K. (2019). *Responsibility and AI*, Council of Europe study. Available at: https://rm.coe.int/responsability-and-ai-en/168097d9c5 (Accessed: 8 February 2024).

[163] *ibid*.

[164] Council of Europe (2023). *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Available at: https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043 (Accessed: 7 February 2024).

1. UPHOLDING RESPONSIBILITY FOR AI OUTPUTS AND DECISIONS

1.1. KEY ACTORS IN THE AI VALUE CHAIN AND THEIR DUTIES

The integration of AI systems in the information and communication space introduces potential risks for human rights violations as outlined in the UN's *Taxonomy of Human Rights Risks Connected to Generative AI*.[165] These include both tangible harms, such as health hazards due to the dissemination of false medical information, and systemic harms, such as biases in AI-generated content that pollute the underlying models and perpetuate inequalities. This integration also challenges access to reliable information within the information ecosystem, as defined in the Principles of the Partnership for Information and Democracy.[166]

Assigning responsibility for mitigating these risks remains paramount to safeguard our democracies, especially given the complex ecosystem of AI development, deployment and use. Yet, there is no standardized method for proportionally distributing responsibility or mechanisms to ensure that AI system developers and deployers take these responsibilities seriously.

Voluntary ethical practices and codes of conduct in the tech industry (discussed in Chapter 3, section 1) acknowledge the need to take responsibility seriously. However, these initiatives often lack institutional mechanisms for external validation, enforcement tools, and strong sanctions, making them insufficient as effective safeguards against the adverse effects of AI.

The need to clearly define these responsibilities and ensure their fulfillment through enforceable legal obligations is grounded in the Universal Declaration of Human Rights (UDHR), which requires States to safeguard individuals and groups from human rights infringements. This responsibility entails adopting preventive measures to ensure the full realization of human rights, particularly those pivotal for maintaining a free, plural, and diverse global information space.[167] Such measures include the establishment of proactive duties for external entities, including AI companies and entities, which are therefore obligated to adhere to legal mandates upholding human rights standards.

The attribution of duties and enforceable legal obligations, deeply ingrained in the UDHR's principles – notably the fundamental rights to justice and effective remedy[168] – is indispensable for guaranteeing accountability in instances of human rights violations by AI systems.

[165] OHCHR (2023). *Taxonomy of Human Rights Risks Connected to Generative AI*. Available at: www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf (Accessed: 7 February 2024).

[166] Forum on Information and Democracy (n.d.), *International Partnership for Information & Democracy*. Available at: https://informationdemocracy.org/international-partnership-on-information-democracy/ (Accessed: 8 February 2024).

[167] United Nations (n.d.). *The Foundation of International Human Rights Law*. Available at: www.un.org/en/about-us/udhr/foundation-of-international-human-rights-law (Accessed: 8 February 2024).

[168] UDHR, Articles 7, 8, and 10.

RECOMMENDATIONS TO **STATES**

- **As a starting point, policymakers should**:
  - **Clearly identify who are the key actors in the AI ecosystem and their responsibilities.**[169] At a minimum, this should encompass AI developers, AI deployers, and users of AI systems for personal, professional or business purposes. This is a crucial step for ensuring that all relevant parties are made aware of their roles and responsibilities.
  - **Ensure that the level of responsibility assigned corresponds to the systemic risks of the AI system to the information space**, i.e., the higher the risks, the greater the requirements to be fulfilled (see Chapter 1, section 2.2, and Chapter 4, section 2).
  - **Track and assign responsibility for both unintended consequences and deliberate misuse or abuse of AI systems**. Such efforts cover the full spectrum of potential issues arising from AI development and deployment to ensure accountability.
- **Once key AI actors and their responsibilities have been identified, policymakers should clearly outline their duties. In doing so, policymakers should**:
  - **Establish multilevel transparency obligations throughout the whole AI value chain** (see section 1.2).
  - **Require adequate documentation and clear communication of risks by AI developers to AI deployers, and from AI deployers to AI users.** This is to ensure the safe application and usage of AI systems. Deployers should then implement guarantees to mitigate risks that may arise during the operation of these systems (see section 2).
  - **Promote mechanisms for prevention and rapid response**. Given the scale and speed at which AI systems operate, it is crucial to establish mechanisms for both prevention of harm and rapid response to any harms that arise (see Chapter 1, section 2.3). This is particularly vital in the context of establishing an accountability system for human rights violations involving AI technologies.[170]
- **In outlining the duties of AI developers and deployers, policymakers should impose a legal obligation to:**
  - **Respect human rights across all business operations and respect ethical principles** such as the ones outlined in UNESCO's *Recommendation on the Ethics of Artificial Intelligence*.[171]
  - **Identify and mitigate human rights risks related to harmful legal and illegal content and use** by implementing best practices, including through:
    - Human labeling for AI training (see Chapter 1, section 1.2); and
    - Red-team activities using state-of-the-art methodologies (see Chapter 1, section 2.1).

[169] For instance, the Digital Services Act (DSA) defines different liability regimes for different "information society services" ("Chapter 2: Liability of Providers of Intermediary Services", Articles 4 - 10).

[170] Yeung, K. (2019). *Responsibility and AI*, Council of Europe study. Available at: https://rm.coe.int/responsability-and-ai-en/168097d9c5 (Accessed: 8 February 2024).

[171] UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

- **Establish robust mechanisms for addressing complaints and feedback from users and other affected parties and remedy human rights harms** they cause (see section 4.3).[172]
- **Adhere to the latest safety and security standards in AI**, continuously updating their practices to reflect state-of-the-art advancements and best practices in the field.
- **Provide transparency about the AI system in a tiered approach**. This means disclosing various aspects of AI systems at different levels, from detailed technical information to regulators and vetted researchers, to operational and impact-related documentation that is comprehensible to the general public[173] (see Chapter 4, section 4.1).

> **In addition, policymakers should impose a legal obligation on AI developers to**:

- **Disclose any limitations**, including restrictions on use, **to deployers and regulators** to prevent misuse, through both documentation and contractual agreements.

> **In addition, policymakers should impose a legal obligation on AI deployers to**:

- **Implement risk mitigation and output moderation measures** in generative AI tools (see Chapter 1, section 2.3).
- **Clearly disclose the capabilities, limitations, and intended use of AI systems** to potential users and the public in the terms of use.
- **Regularly assess AI systems post-release** to identify and mitigate unintended consequences.

> **In outlining the duties of users of AI systems for personal activities, policymakers should encourage:**

- **Informed usage.** Users should be encouraged to proactively seek and review information about the capabilities and limitations of AI systems they interact with.
- **Responsible use of AI systems** in accordance with international human rights law, norms and standards and related domestic legislation.
- **Reporting malfunctions.** Users should be encouraged to report any potential malfunctioning or harmful consequences of AI systems (see section 4.3 and Chapter 4, section 3.1).

> **In outlining the duties of users of AI systems for professional activities, policymakers should impose a legal obligation to:**

- **Adhere to professional ethical standards and codes of conduct** (see Chapter 3, section 1.1).
- **Disclose when using AI and for what purpose.** This would help maintain transparency and accountability, especially in sectors where trust and reliability are paramount.
- **Put in place redress mechanisms for subjects of AI, and feedback mechanisms to report any potential malfunctioning or harmful consequences of AI systems.**

[172] The effectiveness criteria of these mechanisms is outlined in OHCHR (n.d.). *Access to remedy and the technology sector: basic concepts and principles, UN B-Tech Foundational Paper*. Available at: www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf (Accessed: 8 February 2024).

[173] UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

1.2. MULTILEVEL TRANSPARENCY FRAMEWORKS[174,175]

The UNESCO *Recommendations on the Ethics of Artificial Intelligence* emphasize how multilevel transparency (i.e., the disclosure of various aspects of AI systems at different levels, ranging from technical details to broader operational and impact-related information) and explainability are closely linked to ensuring responsibility and accountability.[176]

By facilitating the flow of transparent information throughout the entire system, multilevel transparency can help understand, predict, and possibly prevent harm caused by AI systems.

Moreover, multilevel transparency can play a crucial role in holding responsible parties to account and establishing liability.[177]

RECOMMENDATIONS TO **STATES**

- **Mandate AI developers and deployers to equip AI systems with robust mechanisms for recording information about the operation of the technology** (i.e., through "logging by design" or an equivalent technical solution).[178] **These mechanisms should be fully compliant with applicable data protection and trade secret laws and regulations.** Such a requirement is critical to establish whether and when a risk associated with the technology emerges.[179]

[174] UNESCO (2023). *Multilevel and Meaningful Transparency in Algorithmic Systems: Developing Concrete Criteria to Guide Institutional and Legal Reforms*. Available at: www.unesco.org/en/articles/multilevel-and-meaningful-transparency-algorithmic-systems-developing-concrete-criteria-guide (Accessed: 7 February 2024).

[175] Belli, L. et al. (2022). *Towards meaningful and interoperable transparency for digital platforms, Internet Governance Forum*. Available at: www.intgovforum.org/en/filedepot_download/57/23886 (Accessed: 7 February 2024).

[176] UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

[177] The draft EU AI Act envisions various transparency obligations for AI systems to ensure that their operations are sufficiently clear for users (Article 13).

[178] The draft EU AI Act mandates AI entities to ensure traceability of AI systems through technically allowing AI systems to automatically record events and keep the logs (Article 12, Article 20).

[179] European Union (2019). *Liability for artificial intelligence and other emerging digital technologies*. Available at: https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en (Accessed: 7 February 2024).

2. STRENGTHENING CONTRACTUAL LIABILITY FOR AI SYSTEMS

Absent a regulatory framework for AI liabilities, AI companies have been trying to shield themselves from responsibility and reduce uncertainty with contractual agreements. Nonetheless, risk allocation provisions in contracts are often skewed in favor of AI developers, leading to power imbalances.[180]

To address these ambiguities and ensure that responsible parties are appropriately held accountable in accordance with democratically agreed standards, policymakers and/or the judiciary should clarify the regulatory framework governing these contractual agreements, and their necessary elements. In doing so, they should consider the emerging nature of AI systems across various sectors (including the media sector), the inadequacy of traditional liability models for AI systems, the need for adequate documentation processes, and the importance of transparent risk communication. These considerations are essential to effectively allocate responsibility when something goes wrong.

RECOMMENDATIONS TO **STATES**

> **Clarify that under no circumstances should AI liability be contractually limited or excluded in a manner that violates international human rights law, norms and standards, or that circumvents fundamental legal protections.** This aims to prevent AI developers from abusing contractual agreements to infringe upon basic rights, and from skirting responsibility.

> **Require that contractual agreements between AI developers and AI deployers lay out mutual rights and obligations, including:**

- ♦ **Responsibilities for risk management.**
- ♦ **The attribution of legal responsibility.**
- ♦ **A declaration by AI deployers about their intended use for an AI system prior to obtaining access.**
- ♦ **Provisions to rescind access should the actual usage deviate from the stated purpose.**
- ♦ **Limitations of the system and potential shortcomings.**

> **Require that the terms of use for AI systems specify:**

- ♦ **Liability for AI outputs and decisions.** In general, liability for the outputs and decisions of an AI system lies with its deployer and developer should they fail to comply with required risk-mitigation, transparency, duty of care, and notice and action mechanisms (as specified in section 3.1).[181] They are also responsible for correcting illegal and harmful outcomes. However, users should be held liable in accordance with existing legislation for any harm intentionally caused through the use of AI systems.

[180] Tanenbaum,W., Song, K. and Malek, L. (2022). *Theories of AI liability: It's still about the human element*, Reuters. Available at: www.reuters.com/legal/litigation/theories-ai-liability-its-still-about-human-element-2022-09-20/ (Accessed: 7 February 2024).

[181] The DSA mandates providers of hosting services to allow users to notify them of the presence of information that they consider to be illegal and then review these notices (Article 16).

- **Ownership of data and content.** Users retain ownership of the data and content shared with an AI system. Furthermore, they can opt out of their data and their interactions' metadata being stored and reused for fine-tuning the system. This clarification should not be used by AI deployers to circumvent their obligation to comply with existing copyright and data protection laws. Instead, they should explicitly state how their AI systems adhere to these laws, and ensure a transparent handling of both input and outputs (see Chapter 1, section 1.1.e.).
- **Prohibited use cases for the application of AI systems, in line with international human rights law.**[182]
- **Limitations of the system and potential shortcomings.**
- **The availability and functioning of user feedback** (see Chapter 1, section 2.3, and Chapter 4, section 3.1) **and redress mechanisms** (see section 4.3).[183]

3. IDENTIFYING APPROPRIATE LIABILITY REGIMES FOR AI SYSTEMS

Due to AI system opacity and complexity, there might be cases in which the assignment of liability could be unfair, inefficient or even impossible, or in which an AI subject who suffers harm or damage caused cannot prove causality and ends up without compensation or remedy.[184]

To address these issues, policymakers should reassess existing liability laws and, when appropriate, establish new rules to clarify their applicability to AI systems.

First, it should be specified that in legal actions brought against AI developers and deployers in cases where individuals or a group have suffered damage, the burden of proof should rest on the AI developers and deployers.

Furthermore, in determining the most appropriate liability regime for AI systems in the information and communication space, policymakers should consider several factors. These factors include the functions of AI systems, the nature, severity, probability, and reversibility of the harm caused, and AI developers and deployers' duty of care (see Table 2.1). Conversely, ongoing efforts to differentiate between predictive and generative AI systems, though well intentioned, risk oversimplifying the complex landscape of AI technology, as the lines between these two types of systems are becoming increasingly blurred.

[182] See the prohibited AI practices in the draft EU AI Act (Article 5.1). Additionally, according to the DSA, online platforms shall suspend their services to the users who provide illegal content (Article 23).

[183] The DSA mandates online platforms to provide users with an internal complaint-handling system (Article 20).

[184] European Parliament (2020). *European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence*. Available at: www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.html (Accessed: 8 February 2024).

Table 2.1. Taxonomy of AI systems used in the information and communications space, potential harms and liability regimes

FUNCTIONS	EXAMPLES	POTENTIAL HARMS	LIABILITY CONSIDERATIONS
Content verification and moderation (traditionally done through predictive AI systems)	• Fact checking • Spam detection • Image/video analysis • Hate speech detection • Content flagging	• Bias • Censorship • Missed threats • Privacy violations • Lack of due process	Decisions taken by content verification and moderation algorithms can result in human rights violations. For this reason, AI developers and deployers should be held liable for the harm or damage they cause if they fail to comply with required risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise. Liability should also extend to instances where they fail to establish and effectively manage a clear and predictable mechanism for appealing content moderation actions.
Content curation & recommendation (traditionally done through predictive AI systems)	• Recommendation engines • Personalized news aggregators • Search result ranking algorithms	• Filter bubbles and echo chambers • Polarization • Manipulation	Given the subjective nature of the decisions taken by curation and recommendation algorithms, detrimental outcomes may not necessarily indicate a system malfunction. Therefore, AI developers and deployers should be held liable for failing to comply with required risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise.
Ad targeting & delivery (traditionally done through predictive AI systems)	• Microtargeting • Contextual targeting	• Manipulation • Discrimination • Privacy violations • Deception	Considering that potential harms mainly stem from microtargeting based on protected characteristics and special categories of personal data,[185] AI developers and deployers of algorithms used for microtargeting based on race, color, sex, sexual orientation, gender identity, language, religion, political or other opinion, national or social origin, property, birth or other protected status, as well as special categories of personal data deemed to require higher protection, should be held liable for any resulting harm or damages. This liability should apply irrespective of whether they intended to cause harm or acted negligently[186]. Conversely, AI developers and deployers of algorithms used for behavioral or contextual targeting should be held liable for failing to comply with required risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise.
Content creation (generative AI systems)	• AI writing tools • Image generators • Audio generators • Video generators • AI-driven social media bots	• Mis- and disinformation / deepfakes • Copyright infringements • Privacy and data protection violations • Hate speech and harassment • Illegal speech • Impersonation	While developers and deployers cannot completely eliminate the potential for misuse of generative AI systems, implementing protective measures can significantly reduce the risk of human rights violations. To incentivize adherence to safety and ethical practices, developers and deployers should be held liable if they fail to comply with required risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise. Moreover, AI developers and deployers should be held strictly liable if they do not act upon notices and correct systems accordingly and in a timely manner regarding outputs that are illegal (data protection, copyright, defamation, hate speech, incitement to violence, etc.). Further discussions with civil society, academia, and human rights activists are needed to determine whether a strict liability regime should be applicable in cases of fundamental rights violations. These discussions must also take into account the unique legal and cultural contexts of different jurisdictions.
Personalization (generative AI systems)	• Chatbots • Virtual assistants	• Manipulation • Surveillance/Privacy violations	

Finally, given the significant role platforms that host AI-generated content (such as search engines and social media) play in disseminating and providing visibility to potentially harmful content, it is urgent to establish specific liability mechanisms and risk-mitigation measures.[187]

[185] Special categories of personal data identified by national or regional law as necessitating enhanced protection include data revealing political opinion, trade union membership, health-related data, genetic data, and biometric data (see for example the European GDPR).

[186] Further discussions are needed to define appropriate legislative measures regarding surveillance based advertising which builds on behavioral tracking, as discussed in the OSCE (2021). *Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual*. Available at: www.osce.org/files/f/documents/8/f/510332_1.pdf (Accessed: 8 February 2024).

[187] While the DSA already provides for some elements in this sense, it lacks specific provisions for generative AI as it focuses on the governance of online platforms and search engines.

RECOMMENDATIONS
TO **STATES**

> **Review applicable liability laws or adopt special liability regimes to clarify how they apply to AI systems. In doing so, policymakers should:**
- **Specify that all entities involved in the entire AI value chain, including those who create, maintain, or control the risks associated with the AI system, are in principle liable for AI outputs and decisions.** This approach addresses the challenges of pinpointing specific harmful AI outputs or decisions back to specific human input or design choices, ensuring victims receive compensation or remedy.[188]
- **Specify that the burden of proof should be on AI developers and deployers in legal actions brought against them in cases where individuals or a group have been damaged.**
- **In light of the inherent subjectivity in determining what constitutes "harm", actively engage civil society organizations, academics and researchers across diverse disciplines, backgrounds, and geographies in the process of defining this concept.**
- **Differentiate between the AI system's intended function (i.e., content verification and moderation, content curation and recommendation, ad targeting and delivery, content creation, personalization), the nature, severity, probability, and reversibility of the harm caused, and AI developers and deployers' duty of care within legal frameworks.**

3.1. DEVELOPERS AND DEPLOYERS OF AI SYSTEMS DEPLOYED FOR USE IN THE INFORMATION SPACE

AI systems are playing a growing decision-making role in the moderation, verification, curation, and recommendation of information, as well as in ad-targeting and delivery. These applications vary in function and specific usage, giving rise to different risks. The severity of these risks should be closely linked to the type of liability imposed on AI developers and deployers, as outlined in Table 2.1.

For example, microtargeting users based on race, color, sex, sexual orientation, gender identity, language, religion, political or other opinion, national or social origin, property, birth or other status poses significant risks to fundamental rights. These risks encompass issues such as manipulation, discrimination, privacy violations, and deception. In light of these concerns, AI developers and deployers of AI systems used for ad targeting and delivery based on protected characteristics and special categories of personal data should be held liable for any harm or damage caused by these systems, regardless of the presence or absence of negligence or intentional wrongdoing. Conversely, considering the less intensive invasion of privacy associated with AI systems used for behavioral or contextual targeting, their developers and deployers should be liable for failing to comply with the required risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise, or if it can

[188] European Parliament (2020). *European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence*. Available at: www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.html (Accessed: 8 February 2024). This would also outline that there is a joint and several liability among the different actors in the AI value chain.

be demonstrated that they have engaged in wrongful actions, thereby causing harm or damage to another party.

At the same time, it is important to recognize that in certain situations, despite the potential for human right violations, AI developers and deployers face difficult choices, having to choose between minimizing the risk of false positives or false negatives. For this reason, AI developers and deployers of AI systems used for content verification and moderation should not be subjected to strict liability. However, they should be held liable for failing to comply with required risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise, and for failing to establish and effectively manage a clear and predictable mechanism for appealing content moderation action.[189]

Similar considerations should extend to developers and deployers of AI systems used for content curation and recommendation. In such cases, the subjective nature of the decisions taken by these algorithms may not necessarily indicate a system malfunction. Therefore, developers and deployers should be held liable for failing to comply with required risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise, or if they can be proven to have engaged in wrongful actions resulting in harm or damage to users.

AI systems used for content creation and personalization pose specific and additional challenges to existing liability regimes. These challenges include recognizing that even well-designed protective measures can be circumvented, allocating responsibility when protective measures fail and generative AI systems are exploited for harmful purposes,[190] and establishing a reasonable standard of care for the developers and deployers of these systems.

To incentivize adherence to safety and ethical practices, policymakers should establish that developers and deployers of AI systems used for content creation and personalization should be held liable if they fail to comply with their obligations related to risk mitigation (including output moderation) measures, transparency requirements, and duty of care, unless evidence indicates otherwise. Moreover, developers and deployers of AI systems used for content creation and personalization should face strict liability when failing to effectively manage notice and take-down procedures for illegal outputs, encompassing issues such as data protection, copyright infringements, defamation, hate speech, incitement to violence, and more.

Further discussions with civil society, academia, and human rights defenders are needed to determine whether a strict liability regime should also be applicable in cases of fundamental rights violations. Crucially, these discussions should consider the unique legal and cultural contexts specific to different jurisdictions.

Finally, in jurisdictions where platforms are not typically liable for third-party content, policymakers should specify that AI systems that generate content are not automatically covered by this blanket immunity.

[189] For more information see Forum on Information and Democracy (2022). *Accountability Regimes for Social Networks and their Users*. Available at: https://informationdemocracy.org/wp-content/uploads/2023/01/ID_Report-on-Accountability-regime_Sept22.pdf (Accessed: 8 February 2024).

[190] Henderson, P. (2023). *Who Is Liable When Generative AI Says Something Harmful?* Stanford University. Available at: https://hai.stanford.edu/news/who-liable-when-generative-ai-says-something-harmful (Accessed: 8 February 2024).

RECOMMENDATIONS TO **STATES**

- > **Implement a strict liability regime for developers and deployers of AI systems deployed to microtarget users based on protected characteristics** such as race, color, sex, sexual orientation, gender identity, language, religion, political or other opinions, national or social origin, property, or birth, **as well as special categories of personal data necessitating enhanced protection** (e.g., data revealing political opinion, trade union memberships, health-related data, genetic data, and biometric data). Under this framework, AI developers and deployers should be held liable for any harm or damage caused by these systems, irrespective of their culpability. **Microtargeting based on protected characteristics and special categories of personal data should be banned.**[191]

- > **Implement a fault-based liability regime for developers and deployers of AI systems utilized to target and deliver advertisements based on online behavior (i.e., behavioral advertising) or context (i.e., contextual advertising).** Under this framework, AI developers and deployers should be held liable if they fail to comply with their obligations related to risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise.

- > **Implement a fault-based liability regime for developers and deployers of AI systems used for content verification, moderation and recommendation.** Under this framework, AI developers and deployers would not automatically be held liable for harm resulting from their systems. However, they would be held liable for their failure to comply with obligations related to risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise. Additionally, they would also be liable if they do not establish or maintain a transparent and predictable mechanism for appealing content moderation actions.

- > **Implement a fault-based liability regime for developers and deployers of AI systems used for output generation, such as content or personalization.** Under this framework, AI developers and deployers would be held liable for their failure to comply with obligations related to risk-mitigation measures, transparency requirements, and duty of care, unless evidence proves otherwise. Furthermore, it should be presumed that any harm caused by generative AI systems is due to negligence unless the developers and deployers can demonstrate compliance with all relevant legal and regulatory standards. This approach ensures a higher degree of accountability and encourages stringent adherence to safety and ethical practices in AI development and deployment.

- > **Establish a strict liability regime requiring developers and deployers of AI systems used for output generation, such as content or personalization, to put in place and act upon notice and action procedures for the outputs of their systems.** Under this framework, AI developers and deployers are subject to liability for any harm caused by their systems if they fail to respond promptly to notices and

[191] The DSA regulates microtargeted advertisement in three ways. First, it bans microtargeting (“profiling”) based on the special categories of personal data (Article 26.3). Second, it entirely bans microtargeted advertising when their recipients are minors (Article 28). Third, it mandates online advertisers to provide explanation of the main parameters used in their recommender systems, as well as options for the recipients to modify or influence those main parameters (Article 27). Although beyond the scope of this report, it is recommended that States consider imposing a full ban on microtargeted advertising directed at minors, due to its significant impact on the shaping their freedom of expression and opinion.

correct system outputs that are illegal (including issues related to data protection, copyright, defamation, hate speech, incitement to violence, etc.), without exemptions.

- **Conduct an inclusive dialogue with relevant stakeholders – including, at minimum, civil society and human rights organizations, vulnerable groups, and journalists – to determine in which cases of fundamental rights violations a strict liability regime should apply to developers and deployers of AI systems used for output generation, such as content or personalization.[192] These discussions must also take into account the unique legal and cultural contexts of different jurisdictions.**
- **Clarify by legal means that content generated by AI should not be considered as third-party content or hosting content by the AI system in determining the liability of the generative AI deployer.**

3.2. PLATFORMS HOSTING AI-GENERATED CONTENT AND ENTITIES

One major challenge for platforms, e.g. digital platforms or search engines, hosting AI-generated content and entities is striking a balance between safeguarding freedom of expression and mitigating harm from misleading content, whether human-generated or AI-generated.

To prevent the harmful influence of platforms on the digital public space and promote responsible content moderation and curation, there is a need to define and establish clear responsibilities for the detrimental activities and content they host. This includes AI-enabled detrimental activities such as electoral manipulation, microtargeting, and mass surveillance.

This approach should be based on the presumption that platforms should be liable if there is negligence of risk-mitigation measures, transparency requirements, duty of care, and no implementation of state-of-the-art detection, labeling and authenticity and provenance standards.

RECOMMENDATIONS TO **STATES**

- **Introduce a rebuttable presumption that platforms are liable for illegal content they host (whether human-generated or AI-generated),[193] and the harm they cause unless they can prove that they have:**
  - ♦ **Implemented comprehensive risk-mitigation measures**, including:
    - Easily accessible measures available to all users in order to report illegal content or behavior. This includes timely review and notification of the users concerned and the one who flagged the content, as well as an appeal mechanism.[194]

[192] Such liability could also apply to other cases such as child pornography but goes beyond the scope of this report.

[193] This proposal must be contextualized within the existing legal framework of the United States, particularly in relation to Section 230 of the Communications Decency Act. This section has historically provided broad immunity to online platforms from civil liability for content posted by third parties. Despite numerous challenges and calls for reform, it remains a cornerstone of internet law in the US.

[194] For more information see Forum on Information and Democracy (2022). *Accountability Regimes for Social Networks and their Users*. Available at: https://informationdemocracy.org/wp-content/uploads/2023/01/ID_Report-on-Accountability-regime_Sept22.pdf (Accessed: 8 February 2024).

- Trusted flagger systems[195] and cooperation with independent fact-checkers.[196,197]
- Systemic risk assessments including systemic risks arising from AI-generated content.[198]

♦ **Complied with transparency requirements** (see Chapter 4, section 4)

♦ **Adopted state-of-the-art detection, labeling, provenance and authenticity standards**:

- This includes cooperation with existing standards to maintain metadata or cryptographic signatures to identify provenance and authenticity or watermarks when uploading and sharing content on their platforms.
- Detecting watermarks and provenance and authenticity information and displaying this information to end users.

4. EMPOWERING AI USERS AND SUBJECTS: ENSURING ACCOUNTABILITY IN SYNTHETIC CONTENT CREATION AND MODERATION

Understanding who should bear responsibility for an AI-generated output or decision is vital, but it remains insufficient to ensure true accountability. AI systems, particularly those with significant impacts on individuals or society, require more than just outlining duties or assigning blame.

To achieve true accountability, users and those affected by AI systems must know how these systems operate. Additionally, they need to be informed about when and how AI is used. Finally, they need access to legally enforceable rights in case of adverse impacts caused by AI.

The recommendations provided in this section focus primarily on synthetic content. For a detailed discussion of strategies to empower users impacted by decisions made by AI systems in the information space, particularly in terms of content and account moderation decisions on platforms, readers are encouraged to refer to the Forum on Information and Democracy *Accountability Regimes for Social Networks and their Users* report (2022).[199]

[195] Forum on Information and Democracy (2020). *How to End Infodemics*. Available at: https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf (Accessed: 8 February 2024), pp.27-28 and page 37.

[196] *ibid*, pp. 60, 80-81, 100-101.

[197] The DSA also mandates online platforms to take necessary measures to ensure an efficient operation of trusted flaggers (Article 22).

[198] For a definition of systemic risks see Chapter 1, section 2.2. The DSA mandates online platforms to conduct risk assessment with a focus on human rights (Article 34).

[199] Forum on Information and Democracy (2022). Accountability Regimes for Social Networks and their Users. Available at: https://informationdemocracy.org/wp-content/uploads/2023/01/ID_Report-on-Accountability-regime_Sept22.pdf (Accessed: 8 February 2024).

RECOMMENDATIONS TO **STATES**

- **Develop a comprehensive legal framework that clearly defines the rights of individuals in the context of AI decisions and outputs. This framework should include specific provisions for different types of AI-induced harm, and it should encompass:**
  - **The right of individuals to be informed when an AI system has been used to make decisions impacting them or to create an output concerning them.** This includes a clear explanation of the role and rationale behind the AI involvement in these decisions.
  - **The right to receive explanations of AI decisions and outputs that are technically accurate, yet presented in a manner that is understandable and relevant to the user.** This should include information on the data and criteria used by the AI in making these decisions.
  - **The right to challenge decisions and outputs made by AI, with an assured pathway to a prompt human review.**[200] This process should be easily accessible and designed to provide timely resolutions.
  - **The right to non-discrimination in AI-driven decisions and outputs, ensuring that AI systems do not perpetuate biases or unequal treatment based on race, color, sex, sexual orientation, gender identity, language, religion, political or other opinion, national or social origin, property, birth or other protected characteristics under international human rights law.**

4.1. SYNTHETIC CONTENT AND SYNTHETIC ENTITIES

The integration of AI in the production of text, videos, images, and audio raises significant concerns about transparency and the authentication of content within the information space. Often, users interact with AI-generated content and entities without clear insight into their origins, the actor behind the system, or the technical set-up and functioning of the system (e.g., datasets, training metrics).

To ensure trust and accountability, AI users and subjects should be made aware of when and how AI is being used. This includes fostering transparency about potential biases, limitations, and the accuracy of the AI systems they interact with.

[200] The draft EU AI Act requires human oversight for high-risk AI systems (Article 14).

RECOMMENDATIONS TO **STATES**

> **Mandate AI systems to embed a provenance-indicating watermark in all synthetic content they generate – including text, videos, images, and audio – and to support methods for reliably detecting the content they generate** (see Chapter 1, section 1.5).[201] **Furthermore, mandate that any successive distributors of this content preserve the watermark to ensure transparency about its origin.**

> **Mandate deployers of synthetic entities (e.g., chatbots, virtual assistants) to make users aware that they are engaging with an AI-driven interactive system.**[202]

> **Mandate platforms distributing synthetic content and hosting synthetic entities to adopt a comprehensive policy on the use of synthetic content and accounts, and make that policy available in an easily and clearly understandable manner**. This policy should include:

- ♦ Requirements for labeling AI-generated content, differentiating between (photo) realistic/ authentic-appearing content; content that creates confusion or deceives; AI-generated content used in satire and art; AI-generated content by media organizations, and AI-generated content in political ads and about highly influential actors.
- ♦ Identifying which kind of content is banned and considered illegal and which will receive a warning.
- ♦ Easily accessible flagging systems to flag the synthetic nature of content or accounts.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Make users aware when they are engaging with synthetic entities, such as chatbots or virtual assistants, by clearly labeling the AI's user interface or providing a disclaimer at the beginning of each interaction.**

> **Additionally, complement these disclosure mechanisms with an explicit warning about potential inaccuracies, biases, or falsehoods in the content generated by synthetic entities (i.e., "AI hallucinations"). This will encourage critical evaluation of AI-generated content by users.**

[201] The draft EU AI Act mandates providers of AI systems including GPAI systems that generate synthetic audio, image, video or text content to ensure the outputs of the AI system are marked in a machine readable format and detectable as artificially generated or manipulated (Article 52).

[202] Similarly, the draft EU AI Act envisions that AI systems intended to directly interact with humans should be designed in a way to inform humans that they are interacting with an AI system (Article 52.1). Moreover, providers of AI systems that generate synthetic content shall ensure the outputs of the AI system are marked as artificially generated (Article 57.2).

RECOMMENDATIONS TO **PLATFORMS**

- **Adopt a comprehensive policy on the use of synthetic content and accounts, and make that policy available to users in an easily and clearly understandable manner. This policy should be developed through a participatory and inclusive process which at minimum includes equitable, sustained and substantive participation by independent researchers, civil society and marginalized groups. The process and the selection of stakeholders taking part should be transparent.**
- **Implement a policy where both content and users must acquire a "right of recommendability" before getting promoted or seen in feeds. This right should be granted based on a valid cryptographic signature linked to trusted entities.** Such a policy aligns with the need to carefully select optimization objectives in recommender system development to preserve the integrity of the digital information ecosystem, which is further explored in Chapter 1, section 1.4.
- **Amplify authenticated content in their recommender systems,** particularly of public interest media as certified by the Journalism Trust Initiative.[203]

4.2. USE OF GENERATIVE AI IN POLITICS AND OTHER HIGHLY SENSITIVE SCENARIOS

Political chatbots and AI-generated political content, along with the broader use of generative AI in highly sensitive scenarios (e.g., pandemics, elections, times of conflict), should be subject to the same transparency requirements and limitations as content created by humans.

In light of this, States should reassess and update existing laws – especially election laws – whenever necessary to address the unique challenges of AI in political campaigns and implement appropriate regulation.[204] This includes regulating deepfakes and chatbots used by political parties and governments to engage with voters or disseminate information about electoral processes or political/campaign issues.

In parallel, platforms should continuously reassess their toolkits for combating digital deception, to ensure compliance with the law. This reassessment should focus on evaluating the effectiveness of both policies and technical solutions put in place to tackle the emerging threats posed by sophisticated AI technologies on freedom of expression and opinion, as well as the integrity of the political discourse.

[203] Journalism Trust Initiative. Available at: www.journalismtrustinitiative.org/ (Accessed: 8 February 2024).

[204] The draft DSA guidelines for election integrity elaborated by the European Commission is a step in that direction, European Commission (2024). *Commission is gathering views on draft DSA guidelines for election integrity*. Available at https://digital-strategy.ec.europa.eu/en/news/commission-gathering-views-draft-dsa-guidelines-election-integrity (Accessed on 15 February 2024).

RECOMMENDATIONS TO **STATES**

> **Update election laws to establish clear regulations on the use of generative AI and deepfakes in election campaigns. These regulations should specifically mandate disclosure and transparency requirements regarding the use of such technologies.**

> **Mandate that platforms implement comprehensive risk-mitigation techniques specifically for AI-generated content related to elections and other highly sensitive scenarios.**[205] (see Chapter 1, section 2.3).

> **Mandate tailored red-teaming and risk assessments, and implement strengthened review mechanisms for AI systems (such as chatbots) to be used in political contexts and other highly sensitive scenarios. This recognizes the detrimental effects that hallucinations, or polarizing, biased and discriminatory outputs can have in such settings** (see Chapter 1, sections 2.1 and 2.2)**.**

> **When regulating the use of AI in elections and political processes, it is important to focus on how content is delivered to audiences, rather than the content itself. In this regard, effective measures include banning microtargeting.**[206] (see section 3.1).

RECOMMENDATIONS TO **PLATFORMS**

> **Allocate additional resources to the teams responsible for ensuring safety and respect of human rights,** such as Trust and Safety teams, Human Rights Teams and Content Policy Teams. **Ensure that these teams are fully operational throughout the year, not just during elections or other highly sensitive scenarios.**

> **Continuously reevaluate both policies and technical solutions put in place to combat the risk of digital deception in the realm of political communication and other high-stakes scenarios.** This ongoing review is essential for maintaining legal compliance and ensuring the efficacy of platform strategies in response to the rapidly advancing field of generative AI technologies. Table 2.2 offers some basic guidance, though it is not exhaustive.

[205] For a detailed discussion of policies to be implemented by platforms and mandated by States during elections see "*Protecting Democratic Elections through Safeguarding Information Integrity*" (2024), International IDEA, Forum on Information and Democracy, Democracy Reporting International. Available at: https://informationdemocracy.org/2024/01/30/tech-firms-governments-urged-to-combat-digital-election-threats/ (Accessed 21 February 2024).

[206] As recommended in the Forum on Information and Democracy Report on "*Pluralism of News and Information in Curation and Indexation Algorithms*" (2023) further measures include "prohibit political campaigns and politically affiliated actors from targeting (or intentionally not targeting) lookalike audiences with personalized messages" (p.22).

Table 2.2 Impact of generative AI on standard types of digital deception used in political communication and other high-stake scenarios, legal boundaries, and possible strategies to address them.

Standard types of deception in the realm of political communication and other high-stake scenarios	Use of generative AI in relation to these practices	Legal boundaries[207]	Strategies to mitigate both the adversarial and inadvertent use of gen-AI techniques
Misinformation	Gen-AI systems can unintentionally perpetuate political misinformation, such as conspiracy theories, due to biases in their training data, misunderstandings of context, or AI systems' potential to hallucinate.	The legality of misinformation is a complex issue and intersects with free speech considerations. It also varies based on jurisdiction and the specific nature of misinformation.[208]	• User flagging and appeal mechanisms • Collaboration with trusted flaggers • Viral circuit-breakers • Fact-checking and verification, human and AI technology • Transparency about sources (such as content authenticity and certified media) – collaboration with trusted institutions • Educating the public on identifying and reporting mis- and disinformation and to critically analyze the context of the information they consume • Red-teaming and risk assessments • Address conspiracy theories in reinforcement learning by adding appropriate labels • Reduce the use of engagement optimization for sensitive content including politics and health
Disinformation	Gen AI can help craft and spread disinformation with greater efficiency and believability. Its ability to generate text, images, or videos tailored to specific audiences can make disinformation campaigns more targeted and harder to detect.	Some countries propose or have passed legislation unique to disinformation although these might pose challenges to freedom of speech; for others, proposed amendments or legal basis for tackling disinformation are grounded in other sets of legislation, such as the penal code, civil law, electoral law, laws governing defamation or hate speech, telecommunications law or cybersecurity law.[209]	
Manipulated media/ deepfakes and impersonation	Gen-AI technologies are significantly lowering the cost and easing the process of creating synthetic speeches, statements, or actions attributed to public figures or authorities.	The creation and dissemination of political deepfakes can be subject to legal action under IP law and privacy law. In some jurisdictions (e.g., CA, USA[210] & WA), there are specific laws against deepfakes. Some CSOs in the US argue that deepfakes used in campaign ads are already covered under the law against "fraudulent misrepresentation",[211] but the issue is not settled.[212] Impersonating people online does not immediately classify as a criminal offense,[213] but it could fall under identity theft.[214] Also, it could lead to fraud.[215]	• Use of deepfake technology should be accompanied by clear consent from the individuals being replicated, and content should be clearly labeled as altered or AI-generated. Specific regulations might be needed for satire and art. • Use AI to identify and flag deepfake videos and images. • Cooperate with state-of-the-art detection mechanisms. • Embed invisible markers in authentic media and content to verify originality. • Enforce laws and regulations against the creation and distribution of unlawful deepfakes. • [California and Washington prohibit deepfakes within a certain period of time before an election, unless the communication provides clear and concise disclosure that the deepfake is artificially generated. They also provide exemptions for obvious satire or parody.] • Offer the opt-in opportunity to verify platform users. • Inform users about the risks of impersonation and how to spot it.

[207] The legality of these practices varies significantly across different countries and regions.

[208] Funke, D. and Flamini, D. (2018). *A guide to anti-misinformation actions around the world*, Poynter. Available at: www.poynter.org/ifcn/anti-misinformation-actions/ (Accessed: 8 February 2024).

[209] Bontcheva, K. et al (2020). *Legislative and Regulatory Responses to Disinformation, Excerpt from the Original Report*, Broadband Commission for Sustainable Development. Available at: https://en.unesco.org/sites/default/files/balanceact_legislative_en.pdf (Accessed: 8 February 2024).

[210] Halm, K.C.,Kumar, A., Segal, J. and Kalinowski, C. (2020).*Two New California Laws Tackle Deepfake Videos in Politics and Porn. Available at:* www.dwt.com/blogs/media-law-monitor/2020/02/two-new-california-laws-tackle-deepfake-videos-in (Accessed: 12 February 2024).

[211] Public Citizen (2023). *Comment to FEC: A.I.-Generated Political Deepfakes Are 'Fraudulent Misrepresentation'*. Available at: www.citizen.org/article/comment-to-fec-a-i-generated-political-deepfakes-are-fraudulent-misrepresentation/ (Accessed: 8 February 2024).

[212] Fung, D.O., Brian (2023). *First on CNN: Biden campaign prepares legal fight against election deepfakes*, CNN Politics. Available at: https://edition.cnn.com/2023/11/30/politics/biden-campaign-prepares-against-deepfakes/index.html (Accessed: 8 February 2024).

[213] Bizga, A. (2020). *Blog Post: What is impersonation?,* Bitedefender. Available at: www.bitdefender.com/blog/hotforsecurity/what-is-impersonation/ (Accessed: 8 February 2024).

[214] Bitdefender (n.d.). *What is social media impersonation?*. Available at: https://www.bitdefender.com/cyberpedia/what-is-social-media-impersonation/ (Accessed: 8 February 2024).

[215] Bizga, A. (2020). *Blog Post: What is impersonation?,* Bitedefender. Available at: www.bitdefender.com/blog/hotforsecurity/what-is-impersonation/ (Accessed: 8 February 2024).

Fearmongering and emotional manipulation[216]	AI systems, through the analysis of emotional cues and language patterns, can generate content that is designed to evoke specific emotional responses, such as anger or fear, thereby manipulating public sentiment.	If AI can result in harm or incites violence (or is likely to result in violence), it may be illegal (e.g., anti-LGBTQ fearmongering). However, emotional appeals are commonly used in political communication and campaigns.	• Enforce risk-mitigation measures on platforms to prevent the spread of fear-inducing AI-generated content. • Providing factual information in cooperation with fact-checkers and other independent institutions such as media or administrative authorities to counter fear-based narratives. • Pandemic-specific recommendations to governments:[217] (1) calm, scientifically based messaging from public health authorities; (2) cease and desist warnings directed toward those making extravagant or inappropriate claims; (3) assertive and well-publicized legal action against individuals and entities that make false representations in order to protect consumers rendered vulnerable by their emotional responses to the phenomenon of the pandemic. • Ban targeting content based on sensitive personal data.
Omissions	Generative AI systems can omit certain information either unintentionally or intentionally (e.g., using fine-tuning). This can be particularly concerning if AI systems are used to generate news.	The deliberate omission of critical facts might conflict with laws related to honest communication in advertising and political campaigning.	• Ensure plurality in news reporting. • Human review and editorial responsibility. • Strengthen diversity and representation of AI systems through data provenance, labeling and red-teaming.
Astroturfing	AI can automate the creation of fake profiles and content, simulating grassroots support for a cause or opinion. As gen AI can learn to mimic trends or content patterns, it could make this type of deception tougher to detect.		• Leverage AI to identify and remove bots or fake accounts used in astroturfing campaigns. • Implement regulations that require disclosure of sponsored content and political advertisements. • Amplify authenticated content and users. • Provide user flagging and review mechanisms. • Work with trusted flaggers and fact-checkers.
Influence campaigns targeting specific communities.	AI technologies for style transfer and content adaptation can help craft deceptive and manipulative highly targeted and community-specific influence campaigns. Other than being more effective, these campaigns could also be harder to detect compared with traditional foreign influence operations, especially in non-English speaking regions and Global Majority countries. This capability poses a significant threat to the integrity of elections, not only in terms of misinformation but also in the broader context of shaping political discourse, further disenfranchising already marginalized and underrepresented groups.		• Build infrastructure to monitor the public information space within a country, especially around elections. This includes developing tools to recognize and counter deceptive content tailored for specific communities, with a focus on preserving the integrity of political discourse across diverse linguistic and cultural contexts. • Address information vacuums with reliable information. • Ban targeting content based on sensitive personal data

[216] The AI practices prohibited under the draft EU AI Act includes putting into service an AI system that deploys subliminal techniques, manipulative or deceptive techniques, with the object to affect their behavior (Article 5.1).

[217] Freckleton, I. (2020). *COVID-19: Fear, quackery, false representations and the law*, International Journal of Law and Psychiatry, Volume 72. Available at: https://doi.org/10.1016/j.ijlp.2020.101611.

RECOMMENDATIONS
TO **AI COMPANIES AND ENTITIES**

> **Conduct rigorous and comprehensive red-teaming and risk assessments and implement fastened review mechanisms for AI systems (such as chatbots) to be used in political contexts and other highly sensitive scenarios** (see Chapter 1, section 2).

> **Provide clear and prominent warnings to users about the potential for misinformation that may be inadvertently generated by chatbots. These should include a recommendation to cross-verify information with trusted sources.**

4.3. COMPLAINT HANDLING AND REDRESS PROCEDURES

Individuals and groups might be harmed by AI systems as these can produce illegal and harmful decisions and outputs. To proactively protect and implement user rights during interactions with AI systems and create avenues for compensation and remedy, AI deployers must integrate complaint-handling and redress procedures directly within these AI systems.

Individuals from nations with weak institutions might face unique difficulties in seeking legal recourse in national courts against internationally deployed AI systems. In this context, such procedures become even more crucial. Therefore, organizations deploying AI systems are urged to establish robust mechanisms for lodging complaints and obtaining redress in cases where AI usage breaches rights and legislation regarding, for example, data protection, defamation, or copyright infringement, or when it causes reputational damage.

Of particular concern is the potential reputational harm to media entities from generative AI systems that falsely attribute hallucinated sources, or which fabricate stories appearing to originate from these organizations, pointing to the need to establish a preferential complaint handling and redress process. This scenario is especially problematic in an era where information integrity is paramount, and the line between factual and fabricated news can significantly influence public opinion and discourse.[218]

In this context, national courts play a critical role in arbitrating disputes related to AI-generated content, helping interpret domestic laws in the context of emerging technology and offering guidance on the enforceability of decisions against transnational entities. An AI Ombudsman can also facilitate finding an amicable solution to complaint cases. The interplay between AI systems' internal redress mechanisms, national courts' arbitration roles, and an Ombudsman is further explored in Chapter 4, sections 1.2 and 2.4.

[218] South African Competition Commission (2023). *Final Terms of Reference (ToR) for the Media and Digital Platforms Market Inquiry*, Government Gazette No. 49309. Available at: www.gov.za/sites/default/files/gcis_document/202309/49309gon3880.pdf (Accessed: 8 February 2024).

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Integrate complaint-handling and redress procedures directly within AI systems to enable users to lodge complaints and seek redress in cases where the use of an AI system violates rights and legislation. To guarantee a comprehensive and fair process for addressing grievances, these appeal mechanisms should adhere to the seven principles set forth in the UN's *Guiding Principles on Business and Human Rights* for effective complaint mechanisms. These principles include:**

- **Legitimacy – i.e., the need for these mechanisms to be trusted by the individuals and groups they are intended to serve and to be accountable for conducting grievance processes fairly.** Applying legitimacy to AI systems means ensuring that any complaint-handling and redress mechanism is designed to be trustworthy and impartial. This includes transparency in how decisions are made and ensuring that there are no conflicts of interest that could influence the outcome of a complaint.
- **Accessibility – i.e., the need for these mechanisms to be not only available but also easily understandable and easy to put into effect by all users, regardless of their background and expertise.** This includes making sure that users are aware of these mechanisms, can access them without undue burden, and are provided with the necessary assistance to navigate the process. This might involve user-friendly interfaces, multilingual support, clear instructions, and assurances of safety and confidentiality for those lodging complaints or seeking redress.
- **Predictability – i.e., the need to provide a clear, known procedure with an indicative timeframe for each stage.** In the context of AI systems, applying the principle of predictability means ensuring that users understand how the complaint and redress mechanisms work. This includes clearly outlining the steps involved in lodging a complaint, the expected time it will take to process it, and the possible outcomes or remedies that might result.
- **Equity – i.e., the need to ensure that users who have grievances have reasonable access to sources of information, advice, and expertise necessary to engage in a grievance process on fair, informed, and respectful terms.** In applying this to AI systems, it means that users should have equitable access to the necessary resources to address their concerns. This could involve providing them with clear information about how the AI system works, guidance on the complaint process, and access to expert advice. It also requires ensuring that there are no barriers to effectively engaging with the grievance process, such as complex technical jargon or inaccessible complaint filing procedures.
- **Transparency – i.e., the need to keep parties to a grievance informed about the progress of their case, and to provide sufficient information about the mechanism's performance to build confidence in its effectiveness.** When applied to AI systems, transparency requires clear communication with users about how their specific complaints are being processed and the criteria used to assess them, while respecting confidentiality. Additionally, transparency requires releasing statistics, publishing case studies, or more detailed information on the effectiveness of the mechanism in addressing grievances.

  - **Rights-compatibility – i.e., the need to ensure that the outcomes and remedies provided by the grievance mechanisms are in line with international human rights law, norms, and standards.** In the context of AI systems, applying this principle involves ensuring that the remedies offered for any grievances related to the AI system do not violate human rights principles. Moreover, even when grievances are not initially framed in terms of human rights, the AI system's grievance mechanism should evaluate and address them with consideration for these rights.
  - **A commitment to continuous learning – i.e., the need to regularly analyze and learn from the grievances that arise.** Applying this to AI systems means that the integrated grievance mechanisms should not only address current issues but also evolve based on the insights gained from each case. This involves continuously monitoring and reviewing the types of complaints received, understanding the underlying issues, and making necessary adjustments to the AI system to prevent similar problems in the future. It could also entail updating policies, procedures, or practices based on what is learned from the grievances. Furthermore, these mechanisms should be accessible for individual subjects as well as facilitate collective action when needed.

- **Establish a dedicated channel for complaints lodged by media organizations and human rights defenders. This channel should prioritize issues related to reputational harm caused by the false attribution of fabricated stories or sources that appear to be linked to these media organizations.**
- **Ensure timely review of the complaint, appropriate action, and feedback to the plaintiff and enable referral of the case to a national or international court or an Ombudsman for review if necessary.**
- **Take timely action to address the complaint and correct the system** (see Chapter 1, section 2.3), **which can include deleting personal data from the system and implementing content filters** (see Chapter 1, section 1.1e).

RECOMMENDATIONS TO **STATES**

- **Enact consumer protection laws or other legislation mandating AI systems to establish complaint-handling procedures. These should include the following requirements:**
  - **Timely treatment of complaints prioritizing complaints related to potential human rights abuses.**
  - **A special channel for handling complaints of media representatives and human rights organizations.**
  - **Publication of clear guidelines on the procedures for handling complaints, such as timeframes, eligibility and assessment criteria, expected results and possible remedies.**
  - **Publication of decisions with reasonings and clear references to international human rights law and respective national legislation** (for additional information about redress mechanisms, see Chapter 4, section 2.4).

CHAPTER 3:
INCENTIVIZING ETHICAL AI

INTRODUCTION

Regulations requiring specific measures for ethical AI development, though crucial, represent only one piece of the puzzle. To truly foster a culture of ethical AI, policymakers and other relevant stakeholders must explore alternative approaches that not only discourage harmful practices but also actively incentivize responsible development and use.

One key element in this shift is empowering individuals through AI literacy. By equipping all segments of the population with the knowledge and skills to understand and critically evaluate AI systems, we can create a more informed and engaged public that demands ethical practices, rewards responsible development, and makes responsible use of such tools.

Another element is a supportive ecosystem that incentivizes ethical AI development and deployment through voluntary codes of conduct, certifications, awards, and clear financial incentives such as public procurement and public funds. In addition, this ecosystem must ensure that engineers and those developing AI possess a comprehensive understanding of international human rights laws, norms, and standards.

Finally, it is important to democratize development and access to AI technology and systems to prevent a handful of companies from becoming the gatekeepers of the information space. This includes building public alternatives to for-profit AI systems, ensuring that AI systems can be adapted to the needs and cultural context of countries and communities worldwide.

By fostering AI literacy, providing financial and ethical incentives, and implementing targeted regulations, States can create a future where ethical AI development and use thrive and contribute positively to society as a whole.

1. INCENTIVIZING ETHICAL DEVELOPMENT AND DEPLOYMENT

Ethical considerations in AI are crucial for building an information space that upholds free, pluralistic, and democratic ideals.

To achieve this vision, we must incentivize the development and deployment of AI systems with these values through self-regulatory codes of conduct, robust certifications and rating systems, awards, as well as public funds contingent upon adherence to stringent ethical standards. In addition, we must recognize the potential of open-source and public alternatives to for-profit AI systems, integrate essential ethics-training for IT and AI specialists, establish democratic oversight of AI companies, and provide strong whistleblower protections.

These comprehensive measures work collectively towards the common goal of guaranteeing that the future development of AI systems not only operationalizes democratically established AI priorities, but also safeguards fundamental rights.

1.1 CODES OF CONDUCT

In the absence of comprehensive legislation, self-regulatory codes of conduct have emerged as a significant tool for AI developers and deployers to voluntarily demonstrate their commitment to ethical standards. These codes, such as the G7 Code of Conduct,[219] ideally developed through a participatory process, offer a valuable framework for guiding ethical AI development and deployment. Such codes have also been integrated into legislation guiding companies in their behavior, and are being considered as part of the risk mitigation measures.[220]

However, the key challenge lies in transforming these aspirational documents into truly effective tools. Ensuring robust enforcement mechanisms and addressing implementation hurdles remains crucial for unlocking the full potential of self-regulatory codes and advancing responsible AI practices.

RECOMMENDATIONS
TO **STATES**

> **Recognize and support the role of self-regulatory codes of conduct as a complementary approach to regulation, including through technical assistance, for example, in providing guidance on developing such codes in an inclusive manner.**[221]

> **Develop a system for rewarding companies that adhere to recognized self-regulatory codes.** In this context, **consider linking public procurement opportunities to adherence to recognized self-regulatory codes** (see section 1.4)**.**

[219] European Commission (2023). *Hiroshima Process International Code of Conduct for Advanced AI Systems*. Available at: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Accessed: 8 February 2024).

[220] The DSA mandates the EU Commission to encourage online platforms to draw up voluntary codes of conduct (Article 45), highlighting the code of conduct for online advertising (Article 46) and accessibility (Article 47).

[221] The draft EU AI Act suggests developing Codes of conduct for voluntary application to AI systems other than high-risks AI systems (Article 69).

This can further incentivize commitment and inspire best practices across the industry.

- > **Establish clear channels for AI developers and deployers to report violations of the principles contained in these codes and provide whistleblower protections** (see section 1.8)**.**
- > **Explicitly link adherence to codes of conduct with liability. In this sense, adherence can be considered as one of the required risk management measures** (see Chapter 2, section 1.1)**.**
- > **Require companies to publicly disclose their adherence to codes of conduct, as well as their compliance practices, and support the establishment of mechanisms for scrutiny of these compliance reports by financially supporting independent research and civil society organizations** (see Chapter 4, sections 1.4 and 3.2). To ease scrutiny and foster transparency in evaluating corporate adherence to codes of conduct, States should promote the development of a centralized, open-access, and user-friendly public repository (see Chapter 4, section 4.1).

RECOMMENDATIONS
TO **AI COMPANIES AND ENTITIES**

- > **Adhere to self-regulatory codes of conduct recognized by reputable industry bodies or relevant regulatory authorities.**
- > **Collaborate with diverse stakeholders – including researchers, CSOs, and relevant governmental bodies – to ensure that the principles contained in codes of conduct are comprehensive, meaningful, and consistent with societal expectations, values, and the broader public interest.**
- > **Translate the principles contained in the codes of conduct into practicable steps and integrate them into your development and deployment processes.**
- > **Publicly disclose adherence to self-regulatory codes and demonstrate commitment to ethical AI through regular reporting on company practices and progress. Such reporting should serve to clearly demonstrate to what extent companies comply with commitments, through which measures, and where room for improvement exists.**

1.2 CERTIFICATIONS AND RATINGS

Just as social pressure and public scrutiny drive ethical practices in other sectors, the same can be applied to AI. In this context, a tailored certification system for AI companies and entities, similar to the Fair Trade certification[222] for ethical trade practices, could help incentivize responsible development and deployment. This system would provide a clear and verifiable mark of distinction for companies committed to ethical AI principles, and empower consumers to make informed choices.

While such a certification process would be very comprehensive, a first step to assess the adherence of AI systems to standards could be transparency, safety, or ethical ratings, developed by civil society,

[222] Fair Trade International (n.d.). *About Us*. Available at: www.fairtrade.net/about (Accessed on 9 February 2024).

researchers, or AI authorities. The Stanford University Foundation Model Transparency Index[223] is such an example. Such an index provides useful indicators for users, government, and other actors to be taken into account in procurement rules, and creates healthy competition among AI developers and deployers to improve their systems.

RECOMMENDATIONS TO **CSOS**

Promote the creation of a tailored certification system for AI companies inspired by the success of the Fair Trade certification system. This requires:

> **Developing a comprehensive set of ethical standards for AI development and deployment to anchor the certification.**

> **Establishing an independent, non-profit organization, or building upon an existing organization to manage the certification process.** This body would be responsible for accrediting certification agencies, developing and maintaining assessment criteria, and issuing certifications to qualified companies.

> **Promoting active participation from stakeholders across the AI ecosystem, including developers, deployers, users, researchers, journalists and civil society organizations, with an emphasis on geographic, linguistic, cultural, and cognitive diversity.** This would ensure that a wide range of diverse perspectives, including those of the Global Majority, inform the development and implementation of the certification system.

> Conducting public awareness campaigns to educate users about the importance of ethical AI and the value of the certification system.

> **Establishing an audit mechanism to ensure companies' ongoing compliance with the recommendations.** Companies found to be non-compliant would lose this certification.

> **Establishing a public repository which would allow for public auditing and further increase transparency and trust.**

> **Ensuring financing by requiring AI companies and entities to pay a fee for the certification process.**

RECOMMENDATIONS TO **STATES**

> **Encourage participation in such a "Fair Trade" certification system and a high score in transparency, safety, or ethical ratings by taking these ratings into account in financial decisions impacting AI companies and entities.**

> **Consider making financial resources available to fund the creation and maintenance of such a "Fair Trade" certification system.**

[223] Miller, K. (2023). *Introducing The Foundation Model Transparency Index, Stanford University*. Available at: https://hai.stanford.edu/news/introducing-foundation-model-transparency-index (Accessed: 9 February 2024).

1.3. AWARDS

Award programs at national, regional or international level can be an incentive for ethical AI development and deployment in publicly recognizing such efforts. Such a program needs to rely on clear, fair and public criteria for selection as well as an independent jury. The UN Digital Government Awards are an example of such an approach.[224]

RECOMMENDATIONS
TO **STATES**

> **Put in place an award program to recognize and reward outstanding achievements in ethically aligned AI development and deployment,[225] inclusive of potential tax benefits and/or non-financial rewards such as public recognition.**

1.4. PUBLIC PROCUREMENT AND PUBLIC FUNDS

Public procurement – which represents about 12.9% of GDP in OECD countries,[226] an average of 17% in African nations,[227] 6% in Latin America and the Caribbean,[228] and up to 20% in ASEAN countries[229]– presents a significant opportunity to shape the AI market. By integrating stringent AI ethical standards into procurement processes and prioritizing responsible AI developers and deployers, governments can steer the industry towards ethical practices. This approach empowers governments to not only acquire cutting-edge technologies but also lead by example. In addition, the attribution of public funds and investment should be conditional on AI companies and entities respecting ethical standards, thus ensuring that public funds contribute to developing AI systems that are aligned with the public interest. Tax deductions can be a further measure to incentivize the development and deployment of ethical AI systems.

In championing States' direct involvement in fostering ethical AI development, it is crucial to acknowledge that in some contexts institutions are prone to capture by private interests that can wield significant influence over local policies, including those related to procurement. To safeguard against undue private influence, it is paramount to put robust and transparent public procurement processes in place. This entails transparent and inclusive processes, as well as safeguards in the process such as transparency requirements (for example public disclosure of procurement opportunities, public portals for external actors to verify attribution of contracts and the criteria used to assess attribution) and conflict of interest policies.[230]

[224] UNCTAD (2023). *UN Digital Government Awards celebrate excellence in online public services*. Available at: https://unctad.org/news/un-digital-government-awards-celebrate-excellence-online-public-services (Accessed: 9 February 2024).

[225] This requires the elaboration of clear criteria for such an award program. Awards could be attributed to different AI projects such as those run by local SMEs, non-profit organizations, research institutions or civil society.

[226] OECD (2024). *Size of Public Procurement, Government at a Glance 2023*. Available at www.oecd-ilibrary.org/sites/3d5c5d31-en/1/3/7/1/index.html?itemId=/content/publication/3d5c5d31-en&_csp (Accessed: 9 February 2024).

[227] Arisoy, E., Leipold, K. and Messan, K. (2023). *The expanding role of public procurement in Africa's economic development*, World Bank Blogs. Available at: https://blogs.worldbank.org/governance/expanding-role-public-procurement-africas-economic-development (Accessed: 9 February 2024).

[228] OECD (2020). *Government at a Glance: Latin America and the Caribbean 2020*. Available at: www.oecd.org/publications/government-at-a-glance-latin-america-and-the-caribbean-5ceda53e-en.htm (Accessed: 9 February 2024).

[229] UNDP (n.d.) *Improving Procurement Transparency*. Available at: www.undp.org/asia-pacific/fairbiz/improving-procurement-transparency (Accessed: 9 February 2024).

[230] OECD (2015). *Recommendation of the Council on Public Procurement*. Available at : https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0411 (Accessed: 9 February 2024).

RECOMMENDATIONS
TO **STATES**

> **Develop clear guidelines on what kind of AI systems should be developed, bought, and used in the public sector, and invested in. These guidelines should outline the rules and regulations that the AI systems should respect:**

♦ Whenever possible, favor public alternatives to for-profit systems, open-source systems, or ethical AI systems over non-transparent proprietary solutions.[231]

♦ Whenever possible, favor AI systems certified by the "Fair Trade" certification and scoring high on transparency, safety, or ethical ratings.

> **Use public finances, tax deductions, credits, and exemptions strategically to foster the development and deployment of ethical and local AI systems.**

1.5 OPEN-SOURCE AND PUBLIC ALTERNATIVES TO FOR-PROFIT AI SYSTEMS

As AI systems become increasingly integral to technological advancement and economic growth, access to AI technology becomes crucial for fostering innovation and ensuring equitable participation. Public alternatives to for-profit AI systems and open-source systems are key strategies to address these challenges and build a more inclusive AI future.

Open-source AI (including an open-source license for the code, the data, as well as the system's weights[232] and full release of all three aspects) presents a unique opportunity to democratize access to powerful technologies and foster innovation while promoting transparency and collaboration. This includes enabling building and deploying AI systems held and managed by public entities as well as by civil society, research, media organizations and private actors. However, concerns regarding security vulnerabilities and potential misuse by malicious actors necessitate a careful approach to navigating this complex landscape. Studies have shown that integrated risk reduction or watermarking systems can be easily removed by malicious actors.[233]

Furthermore, States should provide financial resources to support the development of public digital infrastructure and public alternatives to for-profit AI systems, which can counter market concentration and provide access to safe and trusted systems, thus ensuring protection of human rights and respect of ethical standards. The importance of such infrastructure and access for media organizations is also recognized in the Council of Europe's Guidelines.[234] Such public infrastructure includes access to datasets for training of AI systems and computational power to enable a more democratic access for developing AI systems but also the development of AI systems, such as recommender systems or foundational models. Such public alternatives to for-profit AI systems can be administered by independent administrative bodies, institutions similar to public service media, or other actors acting in the public interest such as

[231] The French Digital Law (Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique) encourages the use of open-source software for public administrations developing, buying or using software (Article 16). Available at: www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000033203039 (Accessed: 9 February 2024).

[232] Sijbrandij, S. (2023). *AI weights are not open "source"*, Open Core Ventures. Available at: https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/ (Accessed: 9 February 2024).

[233] Zhang, H. et al. (2023). *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models.* Available at: https://arxiv.org/abs/2311.04378 (Accessed: 15 February 2024).

[234] Council of Europe (2023). *Guidelines on the responsible implementation of artificial intelligence systems in journalism,* Point 6.2 Available at: https://rm.coe.int/cdmsi-2023-014-guidelines-on-the-responsible-implementation-of-artific/1680adb4c6 (Accessed: 7 February 2024).

civil society or researchers. While these systems are preferably published under an open-source license, they can also be kept closed source.

RECOMMENDATIONS TO **STATES**

> **Recognize the dual nature of open-source AI,** acknowledging both its potential for good and its potential for harm. This balanced perspective encourages responsible development and deployment while maximizing the positive impact of this powerful technology. This includes:

- ♦ Ensuring that the same duties, rights, and expected outputs apply to all developers and deployers of AI systems – whether they are open-source, public alternatives to for-profit systems, or proprietary systems – as far as is feasible and legally possible.
- ♦ Engaging in inclusive discussions with relevant stakeholders, including civil society and academia, to develop mechanisms to encourage public interest research and AI development based on open-source systems and harnessing open-source values such as collaboration and transparency. This goal should be pursued in tandem with the establishment of safeguards to protect these systems against malicious use.

> **Establish a dedicated program for developing and maintaining comprehensive public training datasets[235].** These datasets should be broadly representative of the diversity of national populations, high-quality, and readily accessible to researchers and developers globally, particularly in Global Majority countries.

- ♦ **Establish international collaboration frameworks for sharing data and expertise on public training datasets.** This fosters knowledge exchange, facilitates capacity building in under-resourced regions, and promotes the development of global standards for ethical data collection and use.

> **Consider providing public funding to support the development and maintenance of public infrastructure for trustworthy AI systems. This infrastructure includes public alternatives to for-profit recommender,[236] content moderation or classifier systems, and public alternatives to for-profit generative AI systems**. In particular, this could particularly include AI systems used in the information space and by news organizations to democratize access to reliable systems. These systems should preferably be available under an open-source license or an ethical license.[237]

> **Invest in building an alternative digital information space infrastructure which optimizes for societal outcomes, trust, and a democratic and reliable information and communication space.** This can include (providing funding for) building public infrastructure to support the emergence of alternative information and communication spaces, run by community organizations and civil society, or one based on models of public service media. It can also include the provision of funding to conduct research to test how such spaces could function (effects of engagement

[235] In the same spirit, the draft EU AI Act states that the "European health data space will facilitate non-discriminatory access to health data and the training of artificial intelligence algorithms on those datasets" and encourages other sectoral competent authorities to do the same (Point 45).

[236] Forum on Information and Democracy (2023). *Pluralism on News and Information in Curation and Indexing Algorithms*, p.37. Available at: https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf (Accessed: 8 February 2024).

[237] An ethical license includes the four freedoms of open source while putting restrictions on certain behaviors as in Responsible AI Licenses (n.d.). Available at www.licenses.ai/ (Accessed: 8 February 2024)

metrics, authentication schemes and digital identities, encouraging alternative behaviors such as dialogue, etc. See Chapter 1, section 1.4). These systems should preferably be available under an open-source license or an ethical license.[238]

- **Allocate resources to the development, maintenance, and upgrade of essential public digital infrastructure, which may include high-speed internet access, data centers, cloud computing platforms, and adequate computational power.** This will create a foundation for broader AI accessibility and create a demand for AI systems which are responsive to the needs of currently disenfranchised users. **Under no circumstances should these efforts lead to data localization mandates that could potentially jeopardize human rights and/or hinder economic efficiency.**

1.6 ETHICS-TRAINING FOR IT AND AI SPECIALISTS

Education, including university programs, plays a critical role in setting the standards according to which AI system developers will build the systems. To foster ethical development and deployment States must incentivize the integration of modules on ethical standards and the responsibilities of IT specialists in building ethical AI in educational systems. This should also include training on the implications, including potential harmful implications of AI systems for the information space. Such training should be available throughout the whole career of IT and AI specialists.

Moreover, it is essential to acknowledge the broader necessity for subject matter diversity in the education of IT and AI specialists, to develop representative and inclusive AI systems. This implies the need to include human rights laws, cultural studies, history, sociology, and more, to ensure engineers and other specialists involved in AI development and deployment are well-equipped with a comprehensive understanding of the diverse societal impacts their work might have.[239]

AI companies and entities must also invest in training and resources for their own developers, equipping them with the knowledge and tools to recognize and mitigate AI risks. This includes awareness of diverse societal impacts, understanding of human rights laws, norms and standards, identification and correction of biases, and proactive design of systems aligned with ethical principles and democratic values.

RECOMMENDATIONS TO **STATES**

- **Incentivize subject matter diversity in the academic curricula of IT and AI specialists, including human rights law, cultural studies, history, and sociology.**
- **Integrate modules on ethics, ethical standards, the implications of AI systems for the information space, and policymaking in technical training for AI developers** (university courses, specialized schools, etc.) **and offer continuous classes on such issues.** In pursuing this objective, States should:

[238] *ibid*.

[239] Webb, A. (2019). *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*, PublicAffairs New York

- Collaborate with relevant stakeholders, including higher-education institutions, education specialists, civil society, researchers, academics, and other relevant experts, to develop such training programs and university courses.
- Ensure that these modules focus in particular on how AI systems used in the information and communication space may encroach on human rights and fundamental freedoms, including but not limited to:
  - Freedom of expression
  - Media pluralism
  - Representation of diverse cultures and languages
  - Equal access to artistic, scientific, and technological knowledge
  - The right to privacy
  - The right to non-discrimination, encompassing – but not limited to – race, color, gender identity and expression, sexual orientation, language, religion, political or other opinions, national or social origin, property, and birth.
  - Access to information and reliable sources
  - Freedom of the press

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

- **Include AI ethics-training as a fundamental part of the onboarding process for AI developers.**
- **Offer specific continuous training modules for company staff on ethics, ethical standards and the implications of AI systems for the information space.**

1.7 DEMOCRATIC OVERSIGHT OF AI COMPANIES

For most AI companies, board members and investors often decide the strategic directions of the company or entity. Commercial interests (profits and shareholder value) have usually prevailed over public interest and ethical development and deployment of AI systems. To ensure that the public interest guides AI company's strategic decisions, they should be mandated to put in place democratic and participatory governance or oversight structures.

Despites some shortcomings, Meta's Oversight Board[240] is a step in the right direction, as it enables an independent review of content moderation decisions. In the long run, such structures should be mandated to have oversight beyond specific decisions, enabling them to be involved in strategic decisions and policy making of the company.

[240] Oversight Board (n.d.). Available at: www.oversightboard.com/ (Accessed: 9 February 2024).

RECOMMENDATIONS TO **STATES**

> **Mandate AI companies and entities to put in place democratic governance structures.**

- ♦ If the company is not in the country's jurisdiction, licensing processes and public procurement rules can be used to encourage such structures.
- ♦ Provide advice on establishing such mechanisms. These mechanisms can take different forms, such as a supervisory council, citizen assembly, or employer- and user representation.
    - Users should be able to vote for user representatives who can make suggestions, are consulted and have the right to veto decisions that directly affect users such as data protection and data use. These representatives would sit on the board or supervisory council.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Establish democratic governance of the company or entity. Depending on the legal structures and corporate governance practices, this goal could be achieved by:**

- ♦ **Giving members of the public, who are independent of the company and represent the interests of society as well as user representatives, a seat on the board.** This would allow them to have visibility into the company's operations and the authority to scrutinize its activities, ensuring that it operates in the best interests of the public.[241]
- ♦ **Setting up a supervisory council of independent and diverse external experts and user representatives.** The supervisory council should operate as a separate body that would review the company's decisions towards AI systems and provide further guidance on that matter. The council should also have the authority to overturn / issue binding decisions. With their work, the supervisory council should support compliance with international human rights law and promote the ethical and responsible use of AI.[242]

In both cases, AI companies and entities should ensure that the selection criteria for members are fair and transparent, guaranteeing representation from a wide diversity of groups, particularly those most likely to be impacted by the AI system and most vulnerable to it.

[241] Milmo, D. (2023). AI *firms 'should include members of public on boards to protect society', The Guardian*. Available at: www.theguardian.com/technology/2023/dec/06/ai-firms-should-include-members-of-public-on-boards-to-protect-society (Accessed: 9 February 2024).

[242] Kulick, A. (2022). *Meta's Oversight Board and Beyond – Corporations as Interpreters and Adjudicators of International Human Rights Norms,* The Law and Practice of International Courts and Tribunals 2022, Forthcoming. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4226521 (Accessed: 9 February 2024).

1.8 WHISTLEBLOWER PROTECTION

Whistleblowers can provide valuable insights into failures of adhering to mandated governance and accountability frameworks, as well as the malfunctioning of AI systems, by making covered disclosure to the relevant authority. They can participate in investigations or legal proceedings by qualified entities investigating activities, policies, practices, or assigned tasks that are considered to be a potential or suspected violation of existing laws, rules, or regulations.[243] However, current and former employees might hesitate exposing or publicizing such misconduct or malfunctioning due to the fear of retaliation (e.g., direct or indirect dischargement, demotion, suspension, threats, harassment, blacklisting, or any other discriminatory or adverse personnel action)[244] and the lack of incentives and adequate protections.

RECOMMENDATIONS TO **STATES**

- **Set up strong legal protections for whistleblowers who qualify as former or current employees in the AI industry.** These include:
  - **Establishing a mechanism that allows whistleblowers to pursue legal action if they face retaliation as a result of disclosing potential violations of existing laws, rules, and regulations or other potentially unethical behavior. This includes unethical-but-not-illegal disclosures.**[245] This private right of action should include appropriate remedies and protections.
  - **Imposing criminal liability** for the harassment and intimidation of whistleblowers.
  - **Providing measures of physical security for whistleblowers**, such as guaranteeing personal security through safety measures and providing stress and trauma support.
  - **Outlawing confidentiality agreements that prevent or restrict the disclosure of factual information related to potential violations of laws, rules, regulations, ethical standards, or human rights within the realm of AI.**
- **Set up a special platform where whistleblowers can submit their complaints, including anonymously, and have them expeditiously reviewed.** The AI Authority could manage this platform (see Chapter 4, section 1.2).
- **Establish special and confidential channels for legal counseling for whistleblowers, which grants them a right to be represented by lawyers specialized in whistleblower protection.**[246]

[243] Schakowky, J. (2021). *A Bill to provide incentives for and protect whistleblowers under the authority of the Federal Trade Commission, and for other purposes*. Available at: https://schakowsky.house.gov/sites/evo-subsites/schakowsky-evo.house.gov/files/SCHAKO_082_xml.pdf (Accessed: 9 February 2024).

[244] US Department of Labor (n.d.). *Retaliation*. Available at www.whistleblowers.gov/know_your_rights (Accessed: 9 February 2024).

[245] Frances Haugen, a whistleblower from Meta (former Facebook), disclosed behaviors within the company that, while not necessarily illegal, were deemed unethical by many observers.

[246] Marzotto, M. (2023). *Whistleblowers are human rights defenders. So why don't we protect them like they are?* The Signals Network. Available at: https://thesignalsnetwork.org/whistleblowers-are-human-rights-defenders/ (Accessed: 9 February 2024).

RECOMMENDATIONS TO AI COMPANIES AND ENTITIES

- **Set up an internal mechanism where employees can submit complaints, including anonymously, and have them expeditiously reviewed. These complaints should be brought to the immediate attention of top management. Failure to address the issue should result in legal action against them, potentially including claims of negligence under liability law** (see Chapter 2, section 3.1). All the entities that receive information disclosed by the whistleblower should guarantee their confidentiality.
- **Inform employees about the rights of whistleblowers and the legal consequences of any form of retaliation against whistleblowers.**
- **Establish anti-retaliation programs, including appointing an independent officer inside the company, who would review reports and concerns of the employees.**[247]

[247] US Department of Labor (n.d.). *How to Create an Anti-Retaliation Program*. Available at www.whistleblowers.gov/antiretaliation (Accessed: 9 February 2024).

2. INCENTIVIZING RESPONSIBLE USE

AI systems have revolutionized how information is generated, shared, consumed and controlled. From automated reporting to sophisticated data analysis and the personalization of content, AI tools have the potential to enhance journalistic practices, improve the interaction between governments and citizens, and empower marginalized communities.

For example, AI can assist news outlets in tailoring news content to user preferences, analyzing data, and automating basic tasks, freeing up valuable time for in-depth investigative work. Furthermore, it can offer cost-effective tools for political engagement and civil representation, democratizing access and participation. AI can also be harnessed to combat disinformation through automated detection and analysis techniques. Finally, AI tools have the potential to support and enhance creative processes, opening up new avenues for artistic expression.

Incentivizing these positive uses by governments, users, and media should go hand in hand with putting guardrails on irresponsible use practices.

2.1. GOVERNMENTS

AI systems can be widely used to disseminate disinformation and manipulate public opinion, which can be particularly damaging during elections, periods of heightened sociopolitical tension, times of conflict, or of economic tensions. Even if the threats of AI are not fully understood, their potential to do so still erodes the public trust in democratic institutions and processes. To secure the integrity of government information and strengthen public confidence in the democratic process, governments should take proactive measures to incentivize the responsible use of AI systems in content creation, communication and dissemination.

RECOMMENDATIONS TO **STATES**

> **Pursue efforts to establish an international charter or guidelines for the use of AI in government content creation, communication and dissemination. As an initial step, countries or regions can take the lead in developing and adopting this charter. This charter should include:**

- Standards on labeling and watermarking AI-generated or AI-aided content creation and dissemination.
- Commitment to human oversight and measures to reduce risks of discrimination or bias.

> **Organize training modules for government officials to learn about the technical possibilities of AI-related systems, including their abilities and limitations, such as hallucinations and potential copyright infringements.**

> **Adopt content authenticity and provenance standards in all government communication to enhance trust in and integrity of information. For government communication, these standards should include authentication of the author.**

2.2. ART AND SATIRE

For artists and satirists, AI offers more effective and easily accessible ways to create artistic content and satire, yet such creations can appear authentic, thus creating confusion and potential harm among the general public. Mis-and disinformation, deepfakes, and biased and discriminatory outputs can be easily created. This facility necessitates careful consideration of how generative AI, in particular, is used in artistic and satiric content creation, and what labeling or content provenance and authentication should be applied.[248]

RECOMMENDATIONS TO THE ARTISTIC AND SATIRE SECTOR

> **Pursue efforts to establish an international charter or guidelines for the responsible use of AI in artistic and satirical content creation and dissemination, outlining standards on content authenticity and provenance, labeling, watermarking and measures to ensure information integrity. As an initial step, such a charter could be developed and adopted at country or regional level, or by associations representing the sector.**

2.3. MEDIA AND JOURNALISTS

The use of AI in creating synthetic content – whether text, images, video, or audio – by journalists carries significant potential. AI can enhance storytelling, provide richer data visualization, and simulate scenarios for investigative purposes, thus enriching journalistic output.

As trust in media organizations is declining, using AI-generated synthetic content to influence the public without proper disclosure could erode such trust even further. Retaining the public's trust is essential to media organizations' watchdog roles and to their business models.

The use of AI in the context of news creation and dissemination, therefore, needs to be aligned with journalistic ethical standards.

[248] The draft EU AI Act takes a similar approach differentiated transparency requirements for evidently artistic, creative, satirical, fictional analogous work or programme (Article 52.3).

RECOMMENDATIONS TO MEDIA ORGANIZATIONS AND JOURNALISTS

> **Adopt a set of ethical guidelines and professional standards for the use of AI in each media organization or sector-wide.** Such guidelines can be inspired by the Paris Charter on AI and Journalism[249] launched by Reporters Without Borders (RSF) and 16 other organizations. The Charter's ten principles demand, among other things, that journalistic ethics always govern the media's technological choices, that media outlets prioritize human agency and remain accountable, and that AI-driven content personalization and recommendation foster news integrity and diversity. Additionally, it emphasizes that media outlets should draw a clear line between synthetic and authentic content and rely on state-of-the-art traceability and authenticity standards**.** Regarding this last principle, the Partnership on AI's Responsible Practice for Synthetic Media[250] provides useful tactical and technical advice. The Council of Europe's Guidelines on the responsible implementation of artificial intelligence systems in journalism also provide orientation for use of AI by media and journalists.[251]

> **Develop best practices on the use of AI in content creation and dissemination to provide guidance on specific issues and cases. These would include, among others:**

- ♦ **Best practices on the creation of (photo)realistic synthetic content.** Among other things, these should clarify in which cases this content should be used and how it should be labeled.
- ♦ **Best practices on disclosure of AI involvement in content creation,** including when it is necessary to disclose the use of AI tools (e.g., using AI to summarize an article, write a headline, conduct research, analyze data).
- ♦ **Best practices on disclosure of AI involvement in content dissemination, ranking systems of articles, and notifications.**

> **Develop training modules to educate journalists about the advantages and limitations of AI systems.** This should include a particular focus on limitations that might endanger journalistic work such as AI hallucinations, biases, issues of source protection, and copyright of media content (see section 3.2).

[249] RSF (2023). *Paris Charter on AI and Journalism*. Available at: https://rsf.org/en/rsf-and-16-partners-unveil-paris-charter-ai-and-journalism (Accessed: 8 February 2024).

[250] Partnership on AI (n.d.). *PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action*. Available at: https://syntheticmedia.partnershiponai.org/ (Accessed: 9 February 2024).

[251] Council of Europe (2023)*. Guidelines on the responsible implementation of artificial intelligence systems in journalism*. Available at: https://rm.coe.int/cdmsi-2023-014-guidelines-on-the-responsible-implementation-of-artific/1680adb4c6 (Accessed: 7 February 2024).

3. BUILDING AI LITERACY

As AI becomes increasingly integrated into our daily lives, affecting how we interact with information, engage in discourse – and even exercise our basic rights – the need for universal AI literacy becomes crucial. It extends beyond professionals directly working with AI tools, such as journalists involved in news production and distribution, to the public as a whole.

AI literacy programs are essential for both AI users and subjects to develop a clear understanding of how AI systems work, de-mystify the algorithms, and raise awareness about biases that shape their outputs. This includes awareness of the potential to use AI to create deepfakes and mis- and disinformation, to disseminate it widely and to provoke AI-generated approval and sharing of such content. Such awareness reduces potential harm and encourages beneficial impacts. Additionally, understanding the legal and ethical aspects surrounding AI empowers individuals to hold AI developers and deployers accountable and advocate for a fairer and more just information landscape, and to encourage a responsible use of AI by all. Public enforcers, too, need to have a meaningful understanding of these tools, relevant to the effective execution of their public duties. This is why targeted AI literacy programs are vital for safeguarding rights and sustaining democracy in the digital era.

Targeted AI literacy to all can empower individuals to critically assess information, actively contribute to the digital realm, and uphold principles of accountability, diversity, and access to reliable information. This collaborative effort will ensure that everyone has the necessary knowledge to navigate the complexities of AI systems and build a future where they serve as a tool for democratic engagement.

3.1. AI LITERACY AND THE ROLE OF COMPANIES

AI companies and entities bear a significant responsibility in promoting and funding AI literacy.[252] This includes providing easily accessible and understandable information about their systems' functioning, limitations, and potential risks. Users need clear explanations of how these systems work, what data they are trained on and how users' prompts are used, and potential biases these systems may harbor.

By embracing transparency and investing in easily accessible tools to better understand how AI systems work, companies can empower users to engage with AI responsibly.

RECOMMENDATIONS
TO **AI COMPANIES AND ENTITIES**

> **Provide comprehensive, user-friendly explanations of AI systems' functions, limitations, data usage, and potential risks. Offer multiple formats, like infographics and videos, to cater to diverse audiences.**

> **Offer training modules and educational programs targeted at different user groups, including policymakers, and the general public**.

[252] The draft EU AI Act mandates AI entities to enhance AI literacy of their staff (Article 4b).

3.2. AI LITERACY PROGRAMS FOR THE GENERAL PUBLIC, GOVERNMENT, MEDIA, JUDICIARY AND OTHER ACTORS

According to a 2022 UNESCO study to which 51 member countries responded, only 11 have developed and implemented an AI literacy program in education.[253] Adopting an education program on AI in school curricula can thus be seen as an important step to promoting AI literacy among the younger generation.

Beyond AI literacy in the education system, literacy programs for the general public are essential to encourage the ethical use of AI systems and reduce the risks of AI systems to information integrity and trust in democratic institutions. Governments can take inspiration from Finland's 1% AI Scheme, aiming to train 1% of the population based on a free-access university course about the basics of AI systems.[254] Governments in collaboration with universities and civil society organizations should develop dedicated courses on AI system functioning, opportunities, and limitations with a particular focus on the implications for the information space. Collaboration with private companies and continuing education could help to enhance uptake. Special attention needs to be paid to the AI literacy of vulnerable and marginalized groups, for which collaboration with locally implemented organizations is helpful. An example of such an initiative is DataLEADS' FactShala, a media and information literacy program that helps people from small cities and villages across India to assess online information and sift facts from misinformation critically.[255] Specific sectors may need targeted AI literacy training, such as DataLEADS' Global Infodemic Management Course for healthcare workers,[256] which tackles the specific challenge of AI-enabled misinformation in the healthcare sector.

RECOMMENDATIONS TO **STATES**

> **Invest in training and education programs in AI literacy for journalists in collaboration with journalism schools.**

> **Integrate AI literacy modules into existing training programs for public officials, including the judiciary,** providing ongoing opportunities for skill development and knowledge acquisition.

> **Invest in the AI and digital literacy of citizens, especially vulnerable groups and minorities, so that they have technical skills and cognitive capabilities to better navigate through the information spaces, detect mis- and dis-information independently, and use AI systems responsibly.** This can be done through:

- ♦ **Integrating AI literacy into education curricula at all levels of education** (primary, middle, high school and higher education).
- ♦ **Conducting public awareness campaigns addressing both the risks of AI systems** (e.g., bias, AI-enabled manipulation, and mis- and dis-information) **and the potential of AI to democratize information** (e.g., by generating targeted news that is relevant to the specific concerns of users, especially those belonging to underserved communities).

[253] UNESCO (2022). *K-12 AI curricula: A mapping of government-endorsed AI curricula*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380602 (Accessed: 8 February 2024).

[254] Delcker, J. (2019), *Finland's grand AI experiment*, Politico. Available at: www.politico.eu/article/finland-one-percent-ai-artificial-intelligence-courses-learning-training/ (Accessed: 9 February 2024).

[255] DataLeads (n.d.). *FactShala: India's Largest Media Literacy Network*. Available at: https://dataleads.co.in/capacity-building/#FactShala (Accessed: 9 February 2024).

[256] Global Infodemic Management Course (n.d.). *Global Infodemic Management Course for Healthcare Workers*. Available at: https://gimch.org/ (Accessed: 9 February 2024).

- ♦ **Developing globally or nationally available training courses on AI and its implications for the information space**, freely available and widely disseminated. These courses should:
  - Be made available through an online interactive platform or a centralized app with a variety of result-oriented courses that are aimed at developing professional skills and ethical understanding of AI in different age groups.[257]
  - Be multilingual and reflective of the cultural differences of different groups in societies.

> **Support targeted programs for vulnerable groups in developing training material,[258] and provide funding for programs addressing the needs of specific categories of users.[259]**

> **Partner with labor unions and professional associations to integrate AI literacy into professional education in different industries.**

> **Ensure that an effective AI literacy program comprehensively covers various key aspects of AI, including but not limited to topics such as:**

- ♦ Fundamentals of AI (definition, functionalities, common applications, AI value chain)
- ♦ General misconceptions about AI (value neutrality of AI, the risks of AI errors, the concept of AI hallucinations)
- ♦ Bias and discrimination in content creation facilitated by AI
- ♦ Privacy, data protection, and copyright concerns associated with the AI
- ♦ Use of AI in the proliferation of harmful content, defamation, hate speech and illegal speech
- ♦ Use of AI in the creation of misinformation, disinformation, proliferation of social media bots, and microtargeting
- ♦ Tools and skills to detect deepfakes and other forms of synthetic content
- ♦ Understanding AI recommender and content moderation systems
- ♦ Accountability and liability regimes (establishing the responsibility for the harm and non-compliance of AI systems)

3.3. FINANCING OF AI LITERACY

Equitable access to AI literacy requires a multi-pronged approach that combines government funding, global cooperation, and active participation from the private sector, including by contributing a meaningful share of their profits to this policy priority. This means investing in public infrastructure, building strong partnerships across borders, and supporting AI literacy projects, including in Global Majority countries. It is also important to leverage current resources, granting free and open access to information, and supporting grassroots initiatives.

[257] Similar to the free online learning platform AI Campus funded by the German Federal Ministry of Education and Research: https://ki-campus.org/publications

[258] See DataLEADS' FactShala, a media and information literacy program run by DataLEADS with support from the Google News Initiative that helps people from small cities and villages across India to critically assess online information and sift facts from misinformation. Further info available at: https://dataleads.co.in/capacity-building/#FactShala

[259] See DataLEADS' Global Infodemic Management Course for healthcare workers, which tackles the specific challenge of AI-enabled misinformation in the healthcare sector. Further info available at: https://gimch.org/

While artificial intelligence brings considerable potential for improvements in society, as has been discussed throughout the report, AI systems also have negative externalities, for example lowering barriers for bad actors. In a similar vein to environmental taxes, AI companies and entities should bear the costs of such externalities for society. Part of the revenue collected through a tax on AI companies could contribute to an Ethical AI Fund, similar to the Universal Service and Access Funds to which mobile operators and telecommunications providers contribute to enhance internet access worldwide.[260]

RECOMMENDATIONS TO **STATES**

- **Channel additional funding to Global Majority countries through Official Development Assistance (ODA), earmarking these funds specifically for AI literacy programs.**

- **Establish a tax on AI companies and entities to address the societal impact of AI**. A portion of the revenue generated by this tax should be allocated to fund community-based AI literacy programs, public alternatives to for-profit systems, and civil society initiatives (see section 1.5 and Chapter 4, section 3.2). These efforts aim to create a more informed and empowered society while also establishing economic incentives for AI companies and entities to minimize the negative impacts of their AI systems.

  To be effective, this tax should:

  - **Build upon and complement ongoing efforts in collaboration with the OECD to establish a 15% effective minimum tax rate for multinational AI companies in each jurisdiction, irrespective of where they operate.**[261]
  - **Apply to companies and entities deploying AI systems, with targeted exemptions for research, education, and other applications that serve the public interest.**
  - **Be based on the number of users of an AI system, the revenue associated with it, and its risk classification**. This approach would place a greater financial burden on companies and entities deploying larger AI systems with greater potential to cause harm.
  - **Be supported by penalties for non-compliance, including fines and suspension of AI deployments.**
  - **Be developed and implemented in coordination with relevant AI stakeholders, including civil society representatives**, to ensure a strong mandate.

[260] Web Foundation et al; (2018). *Universal Service and Access Funds: An Untapped Resource to Close the Gender Digital Divide*. Available at: https://webfoundation.org/docs/2018/03/Using-USAFs-to-Close-the-Gender-Digital-Divide-in-Africa.pdf (Accessed: 9 February 2024).

[261] OECD (2023). *Outcome Statement on the Two-Pillar Solution to Address the Tax Challenges Arising from the Digitalisation of the Economy*. Available at www.oecd.org/tax/beps/outcome-statement-on-the-two-pillar-solution-to-address-the-tax-challenges-arising-from-the-digitalisation-of-the-economy-july-2023.pdf (Accessed: 9 February 2024).

CHAPTER 4:
AI GOVERNANCE AND OVERSIGHT

INTRODUCTION

As AI systems advance and proliferate at an unprecedented speed and scale, they present unique opportunities and risks for the integrity of information and communication ecosystems. This poses novel challenges and questions for policymakers, who should not only respond to the constantly evolving AI landscape but also anticipate its future implications.

AI systems are increasingly capable of shaping the public policy discourse and swaying public opinion on various issues. AI's swift development, combined with its rapid uptake in use, makes it extremely difficult for policymakers to keep up. This situation is further complicated by insufficient levels of literacy around AI, authorities' lack of adequate financial and human resources, and industry lobbying. These private interests, despite publicly advocating for a state-led regulatory approach, also promote the narrative that regulation stifles innovation and that innovation must inevitably move in the direction they foresee. Additionally, they divert attention from concrete challenges by framing AI risks as a matter of ethics rather than law. Finally, restricted access to proprietary AI systems hinders public interest research, further complicating policymakers' efforts to understand AI systems and address their risks.

While most countries have existing regulations that could be applied to AI challenges, their application often results in significant ambiguities. For example, there is an ongoing discussion about who should own the copyright for AI-generated content, especially when it builds upon pre-existing intellectual property. Additionally, AI systems pose a variety of unique risks to which authorities must respond with particular regulations, such as how to assign responsibility for AI-generated content and decisions. This demands an approach that clearly sets out in law the rights to be protected, including freedom of speech, privacy, data protection, and non-discrimination. It also involves establishing clear rules regarding permitted and forbidden conduct, enforced through accountability regimes.

When regulating AI systems, to mitigate existing and potential harms to people, society, and democratic institutions, policymakers should adopt a balanced approach, tailoring the extent of regulation and duties to the risks AI systems pose. Strict measures can impact innovation and hinder the entry of small-scale companies and startups into the market, thereby reinforcing the dominance of larger AI corporations and entities. Therefore, the most stringent measures should be reserved only for AI systems that can have the most detrimental impact on the information space. At the same time, excessive caution may result in governments over-relying on the goodwill of corporate players and missing out on setting clear democratic rules for the development, deployment, and use of AI systems. Therefore, States need to implement regulatory and governance frameworks that aim to establish protections for reliable, safe, fair, and democratic information and communication spaces, while endorsing an innovative approach towards the development of the AI industry. To this end, a principle-based approach should be adopted, as it provides a comprehensive set of values and objectives that should guide every institutional framework responsible for protecting information ecosystems from the potential harms of AI systems. This approach relies on overarching principles, rather than delving into technical and context-sensitive specifics of how local administrative authorities should be structured.

First, the pursuit of public interest, international human rights, and democratic values must be a central objective for policymakers. In particular, policymaking should consider the needs of diverse populations, incorporating reflections on the sociopolitical implications of AI, especially for vulnerable groups. Policymakers should also put in place effective procedures for the involvement of *civil society organizations, researchers, journalists and other marginalized groups in policymaking, implementation, and oversight.*

Second, policymakers should steer clear of the misleading yet popular philosophical assumption that technologies can be apolitical and value-neutral. Not only do technologies exhibit the values of their developers, deployers, and users, but their direct and indirect sociopolitical implications also transmit certain values.[262]

Third, it is the responsibility of policymakers to challenge the dominant narrative that regulation must come after tech innovations have been able to emerge.[263] *Instead, policymakers should try to be as creative as technology innovators, proactively mapping and responding to both the risks and opportunities presented by AI systems so as to channel innovation in a responsible direction that advances the public interest and strengthens democratic institutions.*

The recommendations endorse the principle of "functional equivalence" enshrined in international regulations. This principle acknowledges that different approaches can achieve the same results within the framework of democratic governance. By allowing for flexibility, functional equivalence effectively ensures due respect for diverse contexts in which AI will be developed, deployed, and used.

To effectively navigate the challenges outlined above, States need to implement effective AI governance with adequate transparency and oversight requirements. These have been recognized, for example, in the Council of Europe's draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law.[264] *Transparency and oversight requirements should encompass several elements, including:*

- *Strong democratic institutions that are responsible for overseeing, developing, and enforcing regulations and accountability regimes.*
- *Democratic oversight, public participation, and equitable, sustained, and substantive civil society involvement.*
- *Conformity assessments that ensure compliance with legal requirements pre-release.*
- *Rigorous auditing mechanisms of AI systems' behavior and their development processes.*
- *A set of tools and mechanisms for users to lodge complaints and seek redress.*
- *Investment in research and access to AI systems for researchers to study systems and hold them accountable.*

To fully realize their potential, these components should align within an international governance framework, thus ensuring a coordinated global approach to AI governance.

[262] Magrani, E. (2019). *New perspectives on ethics and the laws of artificial intelligence.* Internet Policy Review. Available at: https://policyreview.info/articles/analysis/new-perspectives-ethics-and-laws-artificial-intelligence (Accessed: 7 February 2024).

[263] Kretschmer, M. et al (2023). *The risks of risk-based AI regulation: taking liability seriously.* Social Science Research Network. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4622405 (Accessed: 7 February 2024).

[264] Council of Europe (2023), *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Article 7. Available at: https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043 (Accessed: 7 February 2024).

1. ESTABLISHING STRONG DEMOCRATIC INSTITUTIONS

The magnitude of the potential harm AI can inflict on our democracies requires strong and resilient institutions, both existing and new, to meet the challenges of our time. These institutions must be capable of efficiently adopting new laws, enforcing existing laws, and ensuring robust oversight of AI systems, their developers, deployers, and users.

Building strong and independent national institutions with specific AI expertise and clear responsibilities can help governments promote a healthy and trustworthy information environment, strengthen public trust in the information ecosystem, and contribute to the protection of democratic principles in the communication and information space.

1.1. ADDRESSING REGULATORY CHALLENGES POSED BY AI

The rapid development of AI systems has brought focus to the various legal uncertainties and gray areas in existing regulations that can be easily circumvented or abused by private interests driven by commercial, malicious, or other incentives. These uncertainties concern fundamental issues such as data protection, privacy, intellectual property, copyright, and civil liability laws. Furthermore, numerous unanswered questions surround the regulations that have already been enacted (e.g., the Digital Services Act (DSA) and UK Online Safety Act) or are in the process of being drafted or finalized (e.g., the EU AI Act, Brazilian draft AI legislation).

To protect human rights and the democratic information space, States should take a proactive approach to create legal certainty regarding AI. This involves clarifying the applicability of existing legislation to AI systems through consistent interpretation of laws and enacting new legislation and regulations as needed. For example, according to UNCTAD,[265] despite 137 countries worldwide having a data protection law, these are not always aligned with international best practices, potentially rendering them insufficient for addressing AI-related challenges.

In this context, adherence to international best practices although not always sufficient in tackling emerging AI issues can ensure basic compliance with human rights and international law. Furthermore, the implementation of guidelines to complement existing regulations is crucial to ensure their consistent and uniform application to AI.

[265] United Nations Conference on Trade and Development. *Data Protection and Privacy Legislation Worldwide.* Available at: https://unctad.org/page/data-protection-and-privacy-legislation-worldwide (Accessed: 7 February 2024).

RECOMMENDATIONS TO **STATES**

> **Adopt relevant legislation, informed by international guidelines and standards where they exist, to specifically address the issues raised by AI in key legal domains** such as intellectual property, copyright, data protection, privacy, non-discrimination, platform governance, and civil liability.

> **Review existing laws and regulations and evaluate their applicability to the challenges posed by AI systems in the information space, with the aim of identifying gaps and ambiguities.** These laws include, among others, data protection, privacy, intellectual property, copyright, governance of digital platforms, and civil liability.

> **Develop comprehensive guidelines for interpreting and updating existing regulations to effectively address AI-related issues related to the information and communication space. These guidelines should, at minimum:**

- ♦ **Establish clear criteria regarding copyright rights of AI-generated output** especially when it builds upon pre-existing intellectual property.
- ♦ **Clarify the scope of fair use exceptions regarding data curation practices involved in training AI systems**, including how these rules apply to generative AI systems and their outputs.
- ♦ **Specify rights and ownership for metadata generated by AI systems.**
- ♦ **Clarify the application of data protection principles such as the right to be forgotten, consent, and opt-out mechanisms,** and other rights in AI-generated content and systems, enabling AI subjects to manage their digital footprint (see Chapter 1, section 1.e).
- ♦ **Identify gaps and propose amendments to election laws, rules, and regulations** to account for the risks and implications associated with the use of AI-generated content in electoral campaigns (see Chapter 2, section 4.2).
- ♦ **Establish responsibility for AI systems' moderation, verification, curation, recommendation, and ad-targeting and delivery decisions** (see Chapter 2, section 3.1).
- ♦ **Clarify responsibility and liability for AI systems' outputs** (see Chapter 2, section 3.1).
- ♦ **Establish how much autonomy AI systems should have, especially in high-stake scenarios, and clarify the role of human oversight** (see Chapter 2, section 4).
- ♦ **Establish frameworks for ensuring the continuous quality and compliance of AI systems through conformity assessment, and auditing** (see section 2). This includes the need for foundational research in evaluation methodologies aimed at assessing the performance of generative AI systems.

1.2. REGULATORY AND OVERSIGHT AUTHORITIES

Regulations can only realize their objectives if regulatory, oversight, and administrative authorities hold strong institutional independence, the ability to act effectively within their remit, and sufficient resources to enforce and evolve.

Therefore, States should empower existing institutions or create new AI-specific ones. This will ensure that governments maintain expertise in AI-related areas, such as data protection, and human rights. These bodies should be granted new powers to oversee the enforcement of AI regulations and issue implementing acts.

Depending on the national context, these powers might be distributed across different sector-specific authorities or vested into a single authority. In any case, these authorities should have legally guaranteed institutional independence to ensure that they act independently, free from political interference by the government, elected officials, companies, and other actors.[266] Moreover, these authorities should be granted legal power to implement and enforce regulations effectively. This includes granting them investigative powers, such as the ability to request the needed information and documentation of AI systems and to conduct technical studies of the system's functioning. In addition to ensuring independence, the chosen institutional set-up should also guarantee expertise, sufficient resources, adaptability, transparency, accountability, foresight, and proactivity of those authorities.

RECOMMENDATIONS TO **STATES**

> **Establish a new authority or enhance the capabilities of existing authorities** (e.g., data protection authorities, human rights commissions, or competition authorities), **and entrust it/them with the mandate to oversee the enforcement of AI regulations and issue implementing acts and regulations.**[267] This authority should be responsible for monitoring tendencies, identifying evolving or future risks, and proactively educating enforcers and public officials. Its tasks should include:

- ♦ **Establishing a comprehensive regulatory framework for AI development, deployment, and use, aligned with primary legislation.**
- ♦ **Overseeing the implementation of national AI policies and strategies.**
- ♦ **Adopting and enacting implementation acts and regulation.**
- ♦ **Enforcing existing regulations** in coordination with other regulatory authorities.
- ♦ **Imposing fines for non-compliance with relevant AI laws, rules, and regulations,** which may include criminal charges for deliberate misconduct resulting in human rights violations, and the prohibition of AI systems from entering or remaining on the market.
- ♦ **Receiving and handling complaints**, unless an Ombudsman is tasked with complaint handling.
- ♦ **Receiving reports on incidents and recommending remedial action.**
- ♦ **Maintaining a public repository of existing AI systems and their documentation.**

[266] The draft EU AI Act includes various provisions to ensure efficient enforcement of the regulation (Chapter 3: Enforcement, Articles 63-68a).

[267] In the EU, an AI Office, as part of the EU Commission, will be responsible for overseeing compliance with the EU AI Act according to its draft version (Article 55b). also: European Commission (2024). *Commission Decision Establishing the European AI Office.* Available at: https://digital-strategy.ec.europa.eu/en/library/commission-decision-establishing-european-ai-office (Accessed: 7 February 2024).

  - **Issuing and revoking licenses** in jurisdictions where such a regime is put in place.
  - **Carrying out the necessary procedures for the assessment, designation, and notification of third-party conformity assessment bodies and for their monitoring.**
  - **Establishing and publishing criteria** (based on criteria elaborated in Table 1.2) **to categorize AI systems based on the potential risks they pose and their systemic impact on the information space**.
  - **Promoting AI literacy among the general public** in cooperation with private companies, the government and educational actors.
  - **Providing expertise and input to other public institutions** (government, judiciary) as needed.
  - **Fostering international cooperation with international organizations to harmonize AI policies and standards**, with a focus on the detailed compliance work that is likely to involve the creation of and reference to IEEE and ISO standards.

- **Mandate developers and deployers of high-risk AI systems to contribute to the cost of governance by mandating them to pay a supervisory fee.**[268]

- **Establish regulations, safeguards, and commit resources to ensure that the authority(s) in question possess essential characteristics for their effective independent functioning, including:**
  - **Independence** – authorities must operate autonomously from political, commercial, or other external influences that could compromise their objectivity or ability to act in the public interest.
  - **Expertise** – authorities must be staffed by diverse individuals with in-depth knowledge of AI systems and their societal implications, capable of enforcing the law, issuing implementation acts, and providing guidance and opinions. To ensure appropriate staffing of these entities:
    - Put in place robust screening procedures and background checks for hiring staff to avoid conflicts of interest.
    - Mobilize resources to ensure that the staff are equipped with the necessary expertise and state-of-the-art tools to perform their tasks.
  - **Adequate resources** – sufficient funding and qualified staffing are critical to enable these regulatory bodies to perform their oversight functions.
  - **Power to enforce** – authorities need the legal power to implement and enforce regulations effectively.
  - **Adaptability** – given the rapid development of AI systems, authorities must be agile and capable of responding quickly to technological advancements to ensure that regulations remain relevant and effective. They should be proactively monitoring trends, identifying evolving and future risks, and educating enforcers.
  - **Transparency and accountability** – the operations and decision-making processes of regulatory bodies should be open and transparent to build public trust, and ensure accountability. This includes communicating proactively with the various stakeholders, publishing reports on the milestones of their work, and seeking input from the public.

[268] The DSA mandates online platforms to pay an annual supervisory fee (Article 43).

- **Investigative powers** – to be effective, these bodies should be granted comprehensive investigative powers, including full access[269] to algorithms (and their documentation) of AI systems, as well as to the data they train and operate on.

1.3. MULTI-STAKEHOLDER PARTICIPATION IN AI GOVERNANCE

It is essential that a broad range of stakeholders – including the public, civil society, academia, journalists, industry, and advocacy groups – be represented in the AI oversight process to ensure ongoing consideration of diverse perspectives and guarantee continuous democratic scrutiny. To this end, States should create efficient mechanisms for involving different stakeholders in an equitable, sustained, and substantive manner. This can be achieved either by establishing a specific advisory board that advises the AI authority or by integrating different stakeholders directly into the authority's governance structure.

RECOMMENDATIONS TO **STATES**

- **Establish a mechanism to ensure equitable, sustained, and substantive participation of stakeholders, including civil society, researchers, impacted and marginalized communities, and subject matter experts, in the AI regulatory body.** To achieve this goal, **two alternative approaches** can be considered:
  - The formation of an **independent Advisory Board**[270] tasked with providing advice to the AI authority.
  - The **direct integration of stakeholders into the governance structure of the AI authority**.[271]
- **Establish a transparent, inclusive, and accountable mechanism for selecting stakeholders invited to sit on the advisory board or in the AI authority.** The selection process should follow clearly established criteria and aim for a diverse representation of groups and interests.

1.4 RESEARCH CAPABILITIES

AI-based systems are often opaque, leading to significant disparities in understanding of these systems between their developers and deployers and other key stakeholders, such as regulators and policymakers. Independent public interest research plays a crucial role in better understanding how AI systems function and what their implications for the information space are, including potential catastrophic risks. Research is also needed in less economically viable areas, such as on recommender systems that bring positive outcomes for society (see Chapter 1, section 1.4).

Until 2014, the most significant machine-learning models were released by academia. Since then, the industry has taken over. Today, investment in AI and the development of AI models is dominated by

[269] The DSA mandates online platforms to enable third-party auditors access to all relevant data necessary for conducting audits (Article 37.2).

[270] The draft EU AI Act envisions establishing the European Artificial Intelligence Board which will be composed of the representatives of member states (Article 56). Additionally, member states should establish/designate at least one notifying authority and at least one market surveillance authority for the purpose of implementing the AI Act (Article 59).

[271] The draft EU AI Act mandates establishing an advisory forum which will advise the Board and the EU Commission about AI-related issues (Article 58a). Additionally, a scientific panel of independent experts should be established (Article 58b).

the private sector.[272] For example, in 2020, AI investments from the private sector in the EU accounted for 84% and from the public sector only 16%, of which 30% were for Research & Development.[273] This calls for an increased public investment in public-interest research and development.[274]

RECOMMENDATIONS TO **STATES**

- **Establish and fund an independent AI research body, which can be either national or supranational, composed of several independent research laboratories. This AI research body should be responsible for:**
  - Monitoring AI development, researching potential risks, and providing education for enforcers and policymakers.
  - Conducting in-depth causal analyses that go beyond standard audits, thus providing crucial insights into the effects of current AI systems.
  - Evaluating the effectiveness of existing AI regulations in addressing the risks arising from AI development, deployment, and use. This research would provide the basis for policy and regulatory recommendations.
  - Developing alternatives to for-profit AI systems that serve the public interest.
- **Create, maintain, and fund a pool of independent civic experts and researchers to support the enforcement of AI regulations in all their aspects, including monitoring, guidance, and risk assessments.**[275]
- **Put in place competitive and transparent financing schemes to support the emergence of research capabilities in Global Majority countries, including international funds, dedicated Official Development Assistance (ODA), scholarships, and research grants.**
- **Collaborate with the Observatory on Information and Democracy**[276] **to conduct meta-research on AI and build global consensus on contentious topics.**

1.5 NATIONAL AND INTERNATIONAL COURTS

The rapid advances in AI systems also pose major challenges to national and international courts, which are tasked with adjudicating cases related to AI systems and, thus, with interpreting the applicability of existing legislation. Legal professionals, including judges, prosecutors, and attorneys at law need to have the relevant expertise to efficiently represent, review, and enforce the plaintiffs' claims for compensation. They also need to be equipped to respond to appeal cases by AI developers and deployers on decisions taken against them, and to treat criminal offenses by AI system developers and deployers.

[272] Stanford University Human-Centered Artificial Intelligence (2023). *Artificial Intelligence Index Report 2023.* Available at: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Accessed: 7 February 2024).

[273] European Commission (2022). *AI Watch: Estimating AI Investments in the European Union*. Available at: https://ai-watch.ec.europa.eu/publications/ai-watch-estimating-ai-investments-european-union_en (Accessed: 7 February 2024).

[274] Recognising the importance of public interest research and research to effectively support implementation of legislation, the EU has established the European Centre for Algorithmic Transparency to support DSA enforcement. European Commission (2022). *European Centre for Algorithmic Transparency*. Available at: https://algorithmic-transparency.ec.europa.eu/about_en (Accessed: 7 February 2024).

[275] Bengio, Y. (2023). *AI and Catastrophic Risk. Journal of Democracy.* Available at: https://www.journalofdemocracy.org/ai-and-catastrophic-risk/ (Accessed: 7 February 2024).

[276] The Observatory on Information and Democracy is a meta-research project aggregating and synthesizing research to offer periodical and global assessment of the information and communication space. More information: Forum on Information and Democracy. *International Observatory on Information and Democracy*. Available at: https://informationdemocracy.org/mission/ (Accessed: 7 February 2024).

RECOMMENDATIONS TO **STATES**

> **Continuously and substantively train legal professionals, such as judges, prosecutors, and practicing lawyers, about the technical capabilities and legal aspects of AI systems, as well as their implications for the information and communication space.** The training should be designed to not only update legal knowledge but also to provide a sufficient technological understanding of AI to foresee possible risks and opportunities when reviewing cases.

> **Ensure that national and international courts have sufficient financial and administrative resources to hire independent technical experts specializing in AI as needed.**

> **Empower courts and prosecutors with respective investigative powers to mandate AI developers and deployers to disclose information about their AI systems. This includes requesting documentation and inquiring about specific details relevant to the case upon the request of the plaintiff.**[277] Furthermore:

- ♦ If the defendant (AI company or entity) shows reluctance in cooperating with the court and enforcement authorities or refuses to share information, courts should have the power to impose sanctions on the defendant. Additionally, any information not shared by the defendant should be interpreted as if it were shared and favorable for the plaintiff.
- ♦ The power of the court to mandate disclosure should cover not only the trial but also the pre-trial proceedings, thus allowing citizens to decide whether it is worth pursuing the case in the trial phase.

> **Ensure that international courts, including the European Court for Human Rights, the Inter-American Court on Human Rights, the African Court on Human and Peoples' Rights and the International Criminal Court, have enough resources to provide training for their legal professionals and court staff to enhance their understanding of AI systems.** This includes developing educational materials and fostering information sharing among courts in order to ensure a harmonized understanding of AI-related legal issues across participant states.

[277] The draft EU AI act mandates AI entities to cooperate with national authorities and provide them with relevant documentation when needed (Article 23.) The DSA also envisions a similar obligation (Article 10).

2. SETTING UP ROBUST PROCESSES

Conformity assessment, licensing, and auditing are crucial mechanisms to efficiently ensure compliance of AI systems with ethical, legal, and regulatory standards.[278] Together they create a holistic framework of continuous human oversight of AI systems.[279]

Pre-release conformity assessments, either self-conducted or by a third party, act as an ex-ante form of oversight, guaranteeing that AI systems are aligned with existing formal regulations. Licensing processes, also an ex-ante form of oversight, can, when appropriate, add an extra layer of verification by a public authority. This ensures that only legally compliant AI systems are used. Meanwhile, auditing serves as an ex-post form of oversight, assuring continuous compliance of AI systems over time.

Introducing oversight mechanisms into the AI governance framework is imperative, especially considering the vast knowledge collected in other industries on conformity assessments, auditing, and licensing (e.g., finance, data protection, healthcare). This expertise can be efficiently tapped into to establish streamlined and consistent accountability processes for AI systems as well.

To determine which AI systems deployed in the information and communication space should undergo conformity assessments, licensing, and/or auditing processes, they need to be categorized according to their risks and systemic impact on dissemination of illegal content, the exercise of fundamental rights and democratic processes (see Chapter 1, section 2.2). While the taxonomy introduced in Chapter 2 can provide some guidance on the types of AI systems, their classification should be carried out on a case-by-case basis by the relevant AI Authority considering the factors outlined in Chapter 1, section 2.2 and as summarized in the table below.

Table 4.1. Relevant factors determining the risk profile of AI systems impacting the information space

AI System Category	Risks associated with the dissemination of illegal content	Risks impacting the exercise of fundamental rights	Risks impacting democratic processes	Case-by-case factors
AI systems used for content verification and moderation	Yes	Yes	Yes	• Intended purpose • Capability to create harm • Capability to react to harms and correct them • Transparency • Safety and robustness • Accessibility to the general public • Accessibility to bad actors • Number of actual and potential users • Use by critical actors (e.g., media, government) • Type and amount of training data • Capability to act autonomously • Past record of the harm caused
AI systems used for content curation & recommendation	Yes	Yes	Yes	
AI systems used for ad targeting & delivery	Yes (microtargeting)	Yes (microtargeting)	Yes (microtargeting)	
AI systems used to create content	Yes	Yes	Yes	
AI systems used to personalize content	Yes	Yes	Yes	

[278] Laux, J. (2023). Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act. Available at: https://doi.org/10.2139/ssrn.4377481 (Accessed: 7 February 2024).

[279] The draft EU AI Act mandates providers of high-risk AI systems to establish high-quality management systems with the aim of ensuring the overall legal compliance of AI systems with legal regulations during the whole lifecycle of AI systems (Article 17).

In general, the following taxonomy should apply (see Table 4.2):

- Despite the low likelihood of causing harm, low-risk AI systems should undergo an internal conformity assessment to ensure their compliance with democratically established norms, laws, rules, regulations, and principles for example with regards to data protection, privacy, risk reduction in bias, and non-discrimination.
- Medium-risk AI systems should be subject to both ex-ante first-party and third-party conformity assessments, conducted by independent organizations accredited by the State, and ex-post auditing every one to five years, depending on the requirements set-up by the AI Authority.
- High-risk AI systems should be subject to both ex-ante first-party and third-party conformity assessments, conducted by independent organizations accredited by the State, and be subject to annual audits. States, depending on their legal context and taking into consideration potential harms, might also consider establishing licensing processes for high-risk AI systems.
- AI systems designed solely for public interest research purposes, developed by bona fide public interest research institutions and vetted as a public interest project, should undergo a first-party conformity assessment.
- Finally, some AI systems and practices can be considered as prohibited such as microtargeting based on protected characteristics.[280]

This comprehensive approach ensures that all AI systems, regardless of their risk level, are rigorously assessed and monitored, aligning their operation with democratic values and protecting human rights.

Table 4.2. AI system's risk profile and oversight requirements*

	Conformity assessment		Auditing
	First-Party	Third-Party	
Low-Risk	x		
Medium-Risk	x	x	x
High-Risk*	x	x	x

* For high-risk systems, some countries may decide to replace a third-party conformity assessment with a licensing process.

RECOMMENDATIONS TO **STATES**

> **> Mandate the AI authority to establish, publish, and uniformly enforce rules in the classification of AI systems into low, medium, and high-risk categories, and prohibited practices. These rules should undergo periodic revisions to keep pace with technological advancements and their applications in the information and communication space.**

[280] The draft EU AI Act prohibits specific AI practices (Article 5) such as those used for manipulative, exploitative and social control practices.

2.1 CONFORMITY ASSESSMENTS

Pre-release conformity assessment is an ex-ante form of human oversight. Its primary goal is to verify that AI systems comply with recognized technical, ethical, and legal standards. It can be conducted by either AI deployers themselves (i.e., first-party conformity assessment) or independent organizations that have relevant state accreditation (i.e., third-party conformity assessment).

Currently, several high-risk AI systems, such as those integrated into medical devices, are mandated to undergo conformity assessments under existing product safety laws. These assessments serve as a prerequisite for deployment or are required when significant modifications could affect the system's compliance.

Similarly, AI systems designed for the information space, such as those used in social media platforms or for generating news or content, should also be subject to stringent conformity assessments. While all AI systems should conduct first-party conformity assessments, medium-risk AI systems and high-risk AI systems should also undergo third-party conformity assessments conducted by authorized organizations. In both cases, the results of these assessments should be published in a public repository (see section 4.1). This approach aims at striking the right balance between risks coming from AI systems and the potential economic burden on developers and deployers of AI systems. In jurisdictions where high-risk AI systems are mandated to undergo a licensing process, deployers should not be required to also conduct a third-party conformity assessment, in order to prevent unnecessary costs and avoid duplicating efforts.

First-party conformity assessments relieve deployers of additional costs but require them to independently monitor compliance with existing regulations. Third-party assessments of medium- and high-risk AI systems can provide independent verification of an AI system's compliance with standards, thus fostering trust among users and stakeholders.

Overall, conformity assessments should assess the compliance of AI systems with national regulations concerning intellectual property, data protection and privacy rights, civil liability law, AI-specific rules, and cybersecurity standards, etc. Moreover, to ensure fairness and accountability of AI systems, conformity assessments should evaluate AI systems against the principles enshrined in international recommendations such as UNESCO's Recommendation on the Ethics of Artificial Intelligence[281] and the OECD AI Principles.[282]

RECOMMENDATIONS TO **STATES**

- **Mandate first-party conformity assessments for all AI systems to be released within the information and communication space.**
- **Mandate third-party conformity assessments for medium- risk and high-risk AI systems to be released within the information and communication space.**[283]

[281] UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence.* Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

[282] OECD (2019). *Recommendation of the Council on Artificial Intelligence.* Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed on 7 February 2024).

[283] The draft EU AI Act establishes mandatory conformity assessments for high-risk AI systems conducted by state-designated notified bodies (Article 30-32).

> **Ensure that AI developers and deployers provide access to their systems to third-party conformity assessment bodies to effectively conduct the assessment.**

> **Ensure that personnel responsible for conducting conformity assessments, whether first-party or third-party, have relevant expertise in international human rights, norms and standards, and issues pertinent to AI systems intended for use in the information space.**

> **Pursue international instruments to enhance the global acceptance of test results generated by certified third-party conformity assessment bodies (CABs), regardless of their location. These include mutual recognition agreements.[284]**

> **Mandate the publication of the results of first-party and third-party conformity assessments and their sharing with the relevant AI Authority**. They should be published in the public repository (see section 4.1).[285]

> **Direct the AI authority to create a public repository of assessed medium and high-risk AI systems**. **This information should also feed into a supra-national repository maintained by the United Nations.** This international repository would increase global transparency and accountability by enabling AI stakeholders, especially those in countries with weaker institutions, to monitor and compare the conduct and performance of AI developers and deployers across countries.

2.2 LICENSING

Licensing by a competent authority can serve as a powerful tool to proactively minimize potential risks and harms stemming from AI systems by addressing identified issues prior to their public release. It can also compel developers and deployers of AI systems to strive for high-quality standards from the outset.[286]

Nonetheless, licensing comes with significant economic costs, posing a risk of reinforcing the advantageous positions of larger corporations which are more apt to have the resources needed to navigate complex licensing requirements. Licensing can also create an excessive administrative burden and open doors to corrupt practices when the rule of law is not fully established.

Before replacing third-party conformity assessments with a licensing process, States should carefully weigh the potential benefits against the drawbacks. This involves evaluating whether alternative measures are more suitable to achieve the desired outcomes taking into account national specificities and the State's administrative capacity.

To avoid excessive red tape on small companies and preserve competition, States opting for a licensing process should limit its application to high-risk AI systems. Similarly to third-party conformity assessments, AI systems designed exclusively for public interest research developed by bona fide public interest research institutions and vetted as a public interest project should be exempt from licensing requirements to promote innovation in the AI industry and research.

[284] See Articles 38-39 of the draft EU AI Act about the coordination between and recognition of the notified bodies of the other countries.
[285] The draft EU AI Act requires high-risk AI systems to be registered in the EU database before they are put to the market (Article 51, Article 60).
[286] Malgieri, G. and Pasquale, F. (2024). *Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. Computer Law & Security Review,* 52. Available at: https://doi.org/10.1016/j.clsr.2023.105899 (Accessed: 7 February 2024).

RECOMMENDATIONS TO **STATES**

> Consider whether establishing **a licensing scheme is the most effective and appropriate process to assess high-risk AI systems prior to their release. If such a system is chosen:**

- ♦ **High-risk AI systems should be legally required to adhere to a "model of justification and explanation".**[287] This means that to acquire a license, AI deployers should sufficiently explain the intricacies of their AI system, justify the rationale for its release, and prove its compliance with existing laws and ethical guidelines. This includes demonstrating that the system complies with privacy, data protection, non-discrimination, accuracy, accountability, and security requirements.
- ♦ **Assign the responsibility for leading the licensing process to the AI authority. This includes developing guidance, recommendations, and best practices serving as a framework for the developers and deployers of high-risk AI systems during the licensing process.**[288] During the licensing process, the AI authority should:
  - Inquire about the technical and legal soundness of the AI system and request additional information as needed.
  - Independently and first-handedly assess the robustness, security and compliance of high-risk AI systems with democratically established norms, laws, rules, regulations, and principles by testing its features as needed.
  - Pay particular attention to the legality of training datasets, internal checks and balances of AI system deployers, and potential risks of the system to human rights and the information space.
  - Assess the compliance and integrity of the AI system not only through the lenses of legal requirements but also its ethical and societal implications.
- ♦ When granting a license, **regular third-party auditing should be mandated as a condition of keeping the license.** Should the system be substantially modified, deployers need to submit the changes to the AI authority for approval and verification.
- ♦ The **AI authority should create a public registry of licensed high-risk AI systems**.
  - Developers and deployers of high-risk AI systems should be mandated to report AI incidents[289] in this registry, share results of their internal research, and submit risk assessment reports and conclusions of both mandatory and voluntary audits.
  - The AI authority should review this information regularly, assess whether licensed systems comply with the conditions of their licenses and develop statistical formulas and metrics based on which they measure compliance, safety, and fairness of the AI systems.
  - The AI authority should outline specific thresholds, crossing which would result in a warning being issued, temporary suspension of the system, or withdrawal of a license altogether.

[287] *ibid*.

[288] OECD (2022). *Responsible AI licenses: a practical tool for implementing the OECD Principles for Trustworthy AI.* Available at: https://oecd.ai/en/wonk/rails-licenses-trustworthy-ai (Accessed: 7 February 2024).

[289] The draft EU AI Act mandates providers of high-risk AI systems to report incidents to market surveillance authorities in their respective countries (Article 62).

    - Decisions about granting/revoking a license should be subjected to legal appeal if necessary.

- **Exclude AI systems exclusively designed for public interest research purposes, developed by bona fide public interest research institutions, and vetted as public interest projects by a responsible authority[290] from mandatory licensing requirements.** Ensure these systems consistently align with their intended research objectives and prevent any exploitation for commercial purposes through mandatory first-party conformity assessment. This approach must be contingent upon the relevant authority establishing a robust definition of "public interest" that is reasonably resistant to manipulation.

2.3 AUDITING[291]

If not properly regulated and monitored, AI systems have the potential to perpetuate and amplify biases, generate – and possibly amplify – harmful mis- and disinformation, and negatively affect human rights. Their risks can be systemic and have far-reaching implications for the trustworthiness of the information space and democracy. While ex-ante mechanisms of oversight (such as conformity assessment) can ensure initial legal and ethical compliance of AI systems, third-party audits as ex-post oversight mechanisms should be integrated into the holistic AI governance framework to ensure that AI systems continue to align with evolving ethical and legal standards throughout their full lifecycle.

Regular audits ensure continuous oversight and compliance with safety standards, ethical practices, and legal obligations. Third-party audits are common in industries where public safety and trust are paramount, such as cloud computing and cybersecurity.[292] The practices established in these industries can be a valuable source for developing robust and efficient standards for auditing AI.

The practice of auditing AI systems is still evolving, and there is no established consensus over what audits should entail. As discussed in section 4.4, access for researchers to AI systems is crucial in better understanding these systems and establishing guidelines for their auditing. Consequently, to ensure that auditors, developers/deployers of AI systems, and State authorities have a similar understanding of the process, the scope of the auditing process should be clearly defined. Furthermore, the frequency of mandatory audits should reflect the severity, probability, and reversibility of harm of AI systems. While high-risk AI systems should be audited by a third-party auditing annually, medium-risk AI systems should be audited less frequently, such as every one to five years depending on the requirements set-up by the AI Authority.

RECOMMENDATIONS TO **STATES**

- **Subject medium-risk and high-risk AI systems released within the information and communication space to mandatory external auditing. While high-risk AI systems should be subject to an annual audit, medium-risk AI systems should**

[290] The responsible authority could be the AI authority discussed in section 1.2 or the independent vetting body described in section 4.2.

[291] As a mechanism of ex-post oversight, the draft EU AI Act mandates providers of high-risk AI systems to establish a post-market monitoring system, which is based on the post-market monitoring plan (Article 61). The DSA mandates very large online platforms and very large online search engines to conduct third-party audits annually (Article 37).

[292] ISACA (2018). *Auditing Artificial Intelligence.* Available at: https://ec.europa.eu/futurium/en/system/files/ged/auditing-artificial-intelligence.pdf (Accessed: 7 February 2024).

be audited every one to five years, depending on the requirements set-up by the AI Authority**.**

> **Define the scope of the audit of high-risk and medium-risk AI systems. This includes process-oriented and impact-oriented auditing, along with legal and technical assessments**. In this context, develop a detailed checklist that includes all mandatory requirements AI systems should meet.[293] This checklist should include assessments of:

- ♦ Whether personal data is collected, stored, and processed lawfully considering privacy and data protection laws.
- ♦ Compliance with copyright and intellectual property laws.
- ♦ The fairness and explainability of algorithms. This entails investigating causality, detecting opaqueness and bias in the algorithms' decision-making logic, and identifying potential harms they might inflict on specific groups of the population.
- ♦ The adequacy of transparency requirements.
- ♦ The robustness and well-functioning of cybersecurity measures.[294]

> **Equip third-party auditing institutions with the relevant expertise and investigative powers**, including:

- ♦ The ability to request from developers/deployers of AI systems the disclosure of all the relevant information necessary for the assessment.
- ♦ The ability to initiate simulations and monitor AI systems in real-life situations, including direct access to online interfaces.[295]
- ♦ The ability to conduct follow-up interviews with the relevant staff members of the company.

> **In order to ensure their impartiality, mandate third-party auditors to:**

- ♦ Be certified by an official authority.
- ♦ Disclose their funding and conflict of interest before conducting the audit.

> **Develop clear rules and guidance on how the process is structured, what the obligations of the AI developers/deployers and the rights of the auditors are, and what the possible consequences of failing auditing requirements are. This includes:**

- ♦ Mandating developers/deployers of medium-risk and high-risk AI systems to develop detailed documentation and submit it to the auditors.
- ♦ In the post-auditing phase, auditors should provide detailed feedback to the AI developers/deployers, outlining areas that need to be addressed and reasonable timeframes.
- ♦ If AI developers/deployers fail to meet these requirements consistently, they should be subjected to warnings, fines, and temporary/final suspension of the AI system according to the severity of the failure.

[293] *ibid*.

[294] Information Commissioner's Office. *A Guide to ICO Audit Artificial Intelligence (AI) Audits Contents.* Available at: https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf (Accessed: 7 February 2024).

[295] Metaxa, D. et al (2021). *Auditing Algorithms: Understanding Algorithmic Systems from the Outside In*. Available at: https://www.nowpublishers.com/article/Details/HCI-083 (Accessed: 7 February 2024).

RECOMMENDATIONS
TO **AI COMPANIES AND ENTITIES**

> **Ensure the auditability of AI systems by providing detailed, clear, and comprehensive documentation.** This documentation should include:

- Information related to the provenance and curation of training datasets.
- Guidelines for human labeling for AI training and red-teaming.
- The models' architecture and capabilities.
- Robustness testing.[296]
- Risk assessments and their results.
- Conformity assessments and their results.
- Mitigation measures implemented.
- Internal response mechanisms, such as correction, reporting, and complaint mechanisms.

2.4 REDRESS MECHANISMS

Establishing accessible, fair, and efficient private and public redress mechanisms is essential to efficiently enforce legislation. Private redress mechanisms should ensure that companies that fail to comply with legal requirements and inflict harm on individuals or protected groups are mandated to react quickly and pay proportional, fair, and equitable compensations to the victims (see Chapter 2, section 4.3). If plaintiffs are not capable of reaching satisfactory solutions through the AI company's or entity's internal complaint-handling mechanisms, the case should be referred to national courts. If the claimant has exhausted all legal avenues available at the national level, they can escalate their cases to the relevant regional or international courts. Alternatively to seeking a judicial review, an Ombudsman could intervene.

RECOMMENDATIONS
TO **STATES**

> **Establish clear and accessible legal pathways to seek redress**.[297] This should include:

- Defining the criteria based on which courts can decide what constitutes harm (in the information and communication ecosystems).
- Defining what constitutes evidence and what kind of evidence is needed.
- Establishing a system with several pathways for redress, including mandating AI companies to put in place internal complaint-handling mechanisms, enabling the referral to a judicial process and/or an Ombudsman.
- Identifying available schemes for compensation.

> **Provide clear and easily understandable guidance and free legal aid to citizens on these legal pathways.**

[296] Explained as "Robustness testing is a type of testing that is performed to assess the ability of a system or component to function correctly when it is subjected to invalid or unexpected inputs, or when it is operating outside of its specified operating conditions" in GeeksforGeeks, *Robustness Testing*. Available at: https://www.geeksforgeeks.org/robustness-testing /(Accessed: 9 February 2024).

[297] The draft EU AI Act grants users a right to lodge a complaint to the relevant market surveillance authority (Article 68a).

> **Appoint an AI Ombudsman, or strengthen an existing Ombudsman institution** tasked with:
  - Reviewing unsettled complaints to find an amicable solution between the AI company or entity and the plaintiff. This also includes reviewing reports by whistle blowers who are not satisfied with the company's internal review mechanism.
  - If no amicable solution can be found, the Ombudsman can take the case to court, where it will serve as a representative of the plaintiff. [298]

> **Implement measures whereby users can also access collective redress mechanisms**, meaning that when several users suffer from similar harm, they should be able to submit collective complaints and request compensation collectively.

> If AI developers/deployers fail to compensate victims as mandated by the court, they should be subjected to the fines for non-compliance with remedy actions. The amount of these fines should be sufficiently high to incentivize compliance and take into consideration the severity of the non-compliance.[299]

3. ENSURING STAKEHOLDER INVOLVEMENT

Historically, the AI field has seen a disproportionate representation of white and cis-male technologists and corporate interests. This prevalence has come at the expense of voices from a diverse spectrum of human identities, including different races, ethnicities, nationalities, social backgrounds, languages, religions, and various political or other beliefs.[300] As a result, underrepresented and marginalized groups, as well as the wider public, have had limited impact on the direction of AI development. These gaps in representation have led to the deployment of AI systems that are often misaligned with the broader public interest, excessively "Northern" in design, exacerbating existing inequalities and fueling mistrust in AI-powered tools.

To ensure the development of robust, equitable, and socially beneficial AI systems for people everywhere, AI governance mechanisms should offer clear channels for inclusive participation. To this end, States should embrace and promote the principle of epistemological pluralism, which entails the respect of diverse viewpoints, logic, actors, and problems. This is especially relevant considering the fact that AI systems, in their current form, tend to reduce "epistemic biodiversity" (i.e., a rich diversity of knowledge, viewpoints, and ways of understanding and interpreting the world, extending beyond mere ideological differences). Left unaddressed, this trend poses a potential threat to the foundations of democratic pluralism by offering a singular, algorithmically determined perspective that may appear as indisputable objectivity.[301]

[298] Ogunleye, I. (2022). *AI's Redress Problem: Recommendations to Improve Consumer Protection from Artificial Intelligence.* Available at: https://cltc.berkeley.edu/publication/cltc-white-paper-ais-redress-problem/ (Accessed: 7 February 2024).

[299] According to the DSA, the Commission can impose on the online platform fines not exceeding 6 % of their total worldwide annual turnover in the preceding financial year (Article 74). The draft EU AI Act proposes fines between 1 and 7% of the worldwide annual turnover (Article 71)..

[300] United Nations (1948). *Universal Declaration of Human Rights*. Article 2. Available at: www.un.org/en/about-us/universal-declaration-of-human-rights (Accessed: 7 February 2024).

[301] Miller, T. et al. (2008). *Epistemological Pluralism: Reorganizing Interdisciplinary Epistemological Pluralism: Reorganizing Interdisciplinary Research Research*. *Ecology and Society.* 13(2): 46. Available at: https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1034&context=unf_research (Accessed: 7 February 2024).

3.1. STAKEHOLDERS’ PARTICIPATION IN AI GOVERNANCE PROCESSES

In addition to the establishment of an advisory board or integrating stakeholders directly in the AI authority, States should provide proactive measures for civil society, researchers, journalists, and the general public’s involvement in AI governance.

RECOMMENDATIONS TO **STATES**

> **Establish a public portal for citizen feedback on AI systems, regulation and governance, managed by the AI authority, and ensure timely feedback to users submitting comments**.

> **Invite and provide funding to CSOs and independent researchers to participate in third-party conformity assessments** (as detailed in section 2.1), **auditing processes** (as detailed in section 2.3) **and systemic risks assessments** (as detailed in Chapter 1, section 2.2).

> **Mandate AI companies and entities to establish a public notice and complaint portal, where anyone can report harms and risks of the system for the information and communications space.** This is especially relevant when collective rights are violated, providing civil society representatives with the means to engage and ensure accountability of AI systems and their impact on society.

> **Actively seek the input of CSOs, researchers, journalists, and the general public in the development of AI policies including regulation, standardization, and conformity assessment processes,** in alignment with established best practices for stakeholder participation.

> **Establish citizen assemblies, where citizens will deliberate on the risks and opportunities of AI for the information and communication space and the governance options to inform public policies and regulations.** These assemblies should mandatorily include representatives of minority and other vulnerable or marginalized groups.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Establish a public notice and complaint portal where anyone can report the harms and risks of the system for the information space.**

♦ Enable anyone to see complaints that are about the system (and not personal complaints).

♦ Provide an annual report about the complaints, notice, and feedback received including the actions taken to follow up.

3.2 CSO FUNDING

Independent and public interest CSOs play a critical role in ensuring corporate accountability and oversight. Yet, their participation is often taken for granted and not compensated. These organizations need the financial resources to dedicate time and energy to participate effectively and substantively on an ongoing, sustained basis. This is particularly challenging in Global Majority countries, where CSOs may face additional difficulties in securing financing. To support CSOs' continued participation in AI legislative, regulatory, and oversight processes, there is an urgent need to establish funding models that ensure their economic sustainability and maintain their independence from governments and business interests.

RECOMMENDATIONS TO **STATES**

> **Commit to financially compensate CSOs participating in official oversight and AI governance institutions and structures.** Independent selection mechanisms, transparency and accountability mechanisms need to be put in place to guarantee CSOs' independence. This also includes safeguards against conflicts of interest.

> **Establish an independent fund to allocate financial resources for CSOs specialized in AI, democracy and information integrity.** Ensure that the independence of CSOs receiving such funds is guaranteed.

> **Establish mechanisms to make AI companies and entities contribute substantially to CSO funding.** This can include dedicating a specific percentage of the taxation of AI companies (see Chapter 3, section 3.3), and a percentage of fines levied on AI companies.

> **Establish an international fund and dedicate a part of Official Development Assistance (ODA) to fund CSOs specialized in AI, democracy, and information integrity in Global Majority countries.**[302]

[302] OECD (2023). *Official Development Assistance (ODA).* Available at: www.oecd.org/dac/financing-sustainable-development/development-finance-standards/official-development-assistance.htm (Accessed: 7 February 2024).

4. ENSURING TRANSPARENCY AND DATA ACCESS

The current opacity within the AI industry is one of the biggest challenges threatening the democratic information and communication ecosystem. AI systems are often not accompanied by comprehensive documentation, making it difficult to understand their design and development process. Furthermore, there is a lack of transparency in disclosing how AI systems perform and are evaluated, which is critical for assessing their reliability.

This lack of transparency in the AI industry deprives users, watchdogs and regulators of essential insights into AI systems, including the factors impacting them. It also obstructs the ability to verify whether developers and deployers adhere to ethical principles and standards. Additionally, the opacity hampers researchers' capacity to determine causality and contribute to the accumulation of knowledge about the risks and opportunities associated with AI systems. Finally, the opacity also leaves policymakers without the crucial insights required for the formulation of effective regulations.

4.1 DISCLOSURE OF AI SYSTEMS AND DATA

As discussed in Chapter 2, section 1.2, the adoption of a tiered transparency approach is key to addressing the current opacity surrounding AI systems and enabling accountability. This entails ensuring that essential information is accessible to various stakeholders in a way that reflects their unique needs, objectives, functions, and interpretative capabilities.

Within this framework, States play a vital role, establishing minimum transparency requirements for AI companies and entities, their systems, and the data they use.

This section focuses exclusively on transparency requirements for generative AI systems.[303] For a thorough discussion of transparency requirements for platforms and the traditional AI systems they employ, refer to the Forum on Information and Democracy report on *How to End Infodemics* (2020).[304]

RECOMMENDATIONS TO **STATES**

> **Mandate AI developers and deployers to provide transparency about their systems in a tiered approach, providing information to the general public and more detailed information to regulators and vetted researchers.**
Information disclosed to the general public and stored in a public repository should include:[305]

- General information about the algorithms, which includes:

[303] Bell, A. et al (2023). Algorithmic Transparency Playbook. Center for Responsible AI. Available at: https://dataresponsibly.github.io/algorithmic-transparency-playbook/resources/transparency_playbook_camera_ready.pdf (Accessed: 7 February 2024).

[304] Forum on Information and Democracy (2020), *How to End Infodemics*. Available at: https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf (Accessed: 8 February 2024).

[305] The DSA mandates online platforms to regularly publish transparency reports (Article 42). Additionally, the Commission can by simple request require online platforms to provide the information relating to the suspected infringement (Article 67), conduct interviews (Article 68) and carry out inspections on their premises (Article 69).

    - Overview of their functions, main goals and rationales behind their decision-making process, specification of input and output data.
    - Tools of human oversight.
    - Assessment of bias and fairness with a focus on their effect on marginalized groups.[306]
  - General information on AI system capabilities and limitations – its functions, potentials, recommended use cases, and possible harms coming from misuse.
  - General information on the system's controllability – how reliable and aligned a system is with its original purposes; what the likelihood and possible indicators of the unintended harm coming from the system are, and what steps could be taken if the system is out of control.
  - General information about the system impacts – what the studied and possible effects of the system on sociopolitical dynamics and human rights are; what the additional safeguards that could be taken by the consumers to use the product more safely could be.[307]
  - Results of risk assessments.
  - Size, composition, scope, and the extent of human labor used in training of datasets.
  - Data provenance curation practices used in training datasets (see Chapter 1, section 1.a).
  - General technical characteristics of the model, computation capabilities, and its environmental impact.[308]
  - Guidelines used for red-teaming and human labeling for AI training.

- **Information disclosed to regulators and vetted independent researchers should be easily accessible and workable for research purposes, and should be include in a centralized repository** (along with information provided to the general public as described above)**:**
  - Access to "base model" – versions of the model before and after fine-tuning, application programming interfaces (APIs), information on system families, and internals of the system (e.g., metadata).[309]
  - Risk assessment methodologies and results.
  - Detailed information about the training data.
  - Technical documentation, including performance metrics, algorithms and validation processes of AI systems.
  - Findings of the research on the efficiency of risk mitigation measures to protect users from harm and violations of human rights.
  - Findings of internal research as well as the methodologies used to conduct the research.

[306] OECD. *Algorithmic Transparency Recording Standard. Observatory of Public Sector Innovation.* Available at: https://oecd-opsi.org/innovations/algorithmic-transparency-standard/ (Accessed: 7 February 2024).

[307] Anderljung, M. et al (2023). *Towards Publicly Accountable Frontier LLMs.* Available at: https://arxiv.org/pdf/2311.14711.pdf (Accessed 7 February. 2024).

[308] Bommasani, R. et al (2023). *The Foundation Model Transparency Index.* Available at: https://crfm.stanford.edu/fmti/fmti.pdf (Accessed: February 7, 2024).

[309] Anderljung, M. et al (2023). *Towards Publicly Accountable Frontier LLMs.* Available at: https://arxiv.org/pdf/2311.14711.pdf (Accessed: 7 February 2024).

4.2 ACCESS TO DATA FOR INDEPENDENT RESEARCH PURPOSES

Independent research about the impact of AI systems is crucial to enhance their transparency, accountability, and safety. This ultimately ensures their democratic governance.

This requires mandating AI systems' developers and deployers to establish clear pathways for researchers to access data about their training, deployment and use.

However, to simultaneously protect the interests of the public, democratic institutions, user privacy, research independence, and company trade secrets, several safeguards should be put in place. First, States must establish or designate an independent vetting body responsible for reviewing research applications. Second, this body needs to implement and manage a prioritization mechanism to strategically assess and rank research projects based on their scientific merit and alignment with policy priorities and societal interests. Finally, States should mandate clear cybersecurity and privacy safeguards for both developers and deployers of AI systems and vetted independent researchers, thereby ensuring the respect of users' fundamental rights.

For detailed recommendations on researcher access to platforms and their functioning, refer to the Forum on Information and Democracy's report on platform transparency and access to data for research purposes in *How to End Infodemics* (2020).[310]

RECOMMENDATIONS TO **STATES**

> **Mandate AI developers and deployers to grant access to AI system data to vetted independent researchers upon a reasoned request from an independent body** (see 4.1).

> **Establish or designate an independent national body[311] responsible for:**

♦ **Assessing whether a researcher requiring access to AI system data meets the established eligibility criteria and granting them the status of vetted independent researcher.** To be eligible for receiving the status of vetted researcher, a researcher should:

- Be affiliated with a recognized research institution, which includes universities, libraries, research institutes, think tanks, civil society organizations, or any other independent entity whose primary goal is to conduct scientific research or to carry out educational activities on a not-for-profit basis.
- Disclose the funding sources of their research.
- Demonstrate their capability to adhere to data security and confidentiality requirements needed to protect personal data and ensure information integrity.
- Demonstrate that their data access requests and associated timeframes are necessary and proportionate.
- Commit to making their research findings publicly available free of charge.

[310] Forum on Information and Democracy (2020), *How to End Infodemics*. Available at: https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf (Accessed: 8 February 2024).

[311] The DSA already outlines a framework that allows vetted public interest researchers to access data from very large platforms (i.e., online platforms with more than 45 million active users in the EU) under certain conditions. This framework could be adapted to facilitate access to relevant data for research on AI systems. The national Digital Service Coordinators (DSCs) and the European Digital Services Coordinator (EDSC) could be employed to vet research applications related to AI systems deployed within the EU (Articles 40.9-10).

- **Reviewing the research applications submitted by vetted independent researchers to AI companies and entities and, if deemed appropriate, issuing a reasoned request for data access to the AI company or entity. In reviewing received research applications, this body should:**
  - **Assess their scientific merit.** A research application should be approved only if it is in the public interest, and is used for non-commercial purposes. The research applications should preferably be focused on the capabilities, controllability, and impact of AI systems.[312]

    Prioritize research applications according to their feasibility, uniqueness, sociopolitical relevance, and public interest.
- **Mandating developers/deployers of AI systems to establish appropriate cybersecurity and privacy safeguards with which vetted independent researchers should comply when receiving, handling, or analyzing qualified data.**
- **Offering vetted independent researchers' advice about necessary privacy and cybersecurity measures and assisting them with complying with the cybersecurity and privacy standards.**

RECOMMENDATIONS TO **AI DEVELOPERS AND DEPLOYERS**

> **Enable vetted independent researchers to access AI data for conducting research in the public interest** (see section 4.1) **and establish mechanisms for providing this access in a manner that does not compromise data privacy, intellectual property, and security.**

> **Notify users that their data is shared for research purposes through posting notices or other appropriate means, and keep them informed of their privacy protections and the type of information shared.** Users whose profile does not host public content should be given an opportunity to opt out of their information being shared with a vetted independent researcher.

4.3 EXPERIMENTAL EVALUATIONS ON PLATFORMS

Undoubtedly, A/B testing (i.e., the method of experimental evaluation aimed at comparing two similar versions of something to compare their performance in a controlled manner)[313] is the most effective methodology for establishing causality when studying the impact of AI systems on user behavior. Compared with methodologies that rely on observing data, A/B studies provide a powerful mechanism to ascertain the direct impact of these systems.

[312] Anderljung, M. et al(2023). Towards Publicly Accountable Frontier LLMs. Available at: https://arxiv.org/pdf/2311.14711.pdf (Accessed: 7 February. 2024).

[313] Gallo, A. (2017). *A Refresher on A/B Testing.* Harvard Business Review. Available at: https://hbr.org/2017/06/a-refresher-on-ab-testing (Accessed: 7 February 2024).

Since A/B testing capabilities are typically proprietary and restricted to companies owning the system, collaboration between external researchers and the private sector is crucial.

Currently, no jurisdiction has a legal mandate for AI companies and entities to allow external vetted independent researchers to conduct this type of experimental intervention.[314]

While a few examples of voluntary collaboration between AI companies and entities and independent researchers exist,[315] States should establish a legal mandate to ensure that experiments in the public interest can be conducted independently of the discretion of tech executives.

RECOMMENDATIONS TO **STATES**

> **Mandate AI developers and deployers to grant vetted independent researchers the possibility of conducting experimental evaluations of AI systems upon reasoned request from an independent body.**

> **Instruct the independent (inter)national body discussed in section 4.2 to develop eligibility and prioritization criteria for conducting experimental evaluations of AI systems in cooperation with academia, civil society representatives, and other relevant AI stakeholders.**

> **Establish a legal requirement for vetted independent researchers to sign a non-disclosure agreement regarding users' confidential information and companies' competitive advantage.**

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Enable vetted independent researchers to conduct experimental evaluations of AI systems in the public interest and establish clear mechanisms for conducting such evaluations in a manner that does not compromise data privacy, intellectual property, and security.**

[314] The potential inclusion of A/B tests in the scope of powers to be granted by external researchers under the DSA remains uncertain. This aspect is currently under discussion and subject to ongoing advocacy efforts. The delegated regulation that will finalize these details is still in development. However, it is important to note that the already finalized delegated act on auditing under the DSA does grant auditors the authority to test recommender systems using appropriate methods, as outlined in clauses 25-30 of the document. European Commission (2023). *Delegated Regulation on independent audits under the Digital Services Act.* Available at: https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act (Accessed: 7 February 2024).

[315] The Toronto/Berkeley Meta Newsfeed project, a collaboration between Meta and academic researchers from Berkeley's Center for Human-Compatible AI and the University of Toronto's Schwartz Reisman Institute and Vector Institute, exemplifies voluntary collaboration between an AI company and independent researchers. Co-led by Jonathan Stray and Gillian Hadfield, the project aims to optimize Facebook's newsfeed recommender algorithm for measures other than user engagement. As summarized in: GPAI (2022). *Transparency Mechanisms for Social Media Recommender Algorithms: from Proposals to Actions. Report,* Global Partnership on AI. Available at: https://gpai.ai/projects/responsible-ai/transparency-mechanisms-for-social-media-recommender-algorithms.pdf (Accessed on 8 February 2024).

4.4 ACCOUNTABILITY SANDBOXES FOR AI ALGORITHMS

Although some jurisdictions (notably the EU) now require platforms to share documentation about their algorithms and provide access to data for researchers, companies tend to be reluctant when it comes to sharing in-depth information such as the source code of their algorithms with external researchers. Their hesitancy is often based on legitimate grounds, such as the need to protect trade secrets.

To facilitate accountability while addressing companies' concerns, States should mandate the establishment of "accountability sandboxes". Unlike conventional sandboxes used for technology development, accountability sandboxes would grant external stakeholders the access to test algorithms and AI systems. They could input data into the AI system in the sandbox and gain insights into an AI system's functioning.

As a result, these new testing environments would allow independent researchers, civil society organizations, and regulatory authorities to hold AI companies and entities accountable for the design of their AI systems, all while safeguarding the confidentiality of the underlying code and parameters.

RECOMMENDATIONS TO **STATES**

> **Mandate AI companies and entities to establish an "accountability sandbox" accessible to external stakeholders,** including independent researchers, civil society organizations, and regulatory authorities. **This "accountability sandbox" should enable access to the AI systems they develop and/or deploy through an application programming interface (API)**, allowing external and independent testing and analysis of these systems while safeguarding proprietary information.

RECOMMENDATIONS TO **AI COMPANIES AND ENTITIES**

> **Set up an "accountability sandbox" providing access to AI systems developed and/or deployed via an API for external and independent testing and analysis.**

5. PROMOTING INTERNATIONAL COOPERATION AND REGULATION

Global governance of AI through international cooperation and regulation can play a crucial role in shaping the AI landscape.

Such governance can help establish global standards, principles, and potentially rules that must be considered in the development, deployment, and use of AI systems in the information space. These norms and regulations should be crafted in the interest of democracy, information integrity, and upholding human rights of citizens worldwide.

International cooperation can also promote effective AI governance at the national level through peer learning, the sharing of best practices, and international leverage. Moreover, it can support capacity and knowledge building, particularly in countries with less-resourced institutions.

In this context, addressing the unique challenges and opportunities of Global Majority countries is essential for ensuring responsible and ethical AI practices. This includes redefining policy targets about fairness to address issues specific to Global Majority countries, such as caste discrimination. Also, as data privacy regimes in the Global Majority countries are less developed compared to OECD levels, international cooperation can help countries establish robust privacy regimes upon which AI regulations can be built.

Finally, international cooperation can enhance the capabilities of Global Majority countries. This includes building their technical capacities and fostering an understanding of the impact of AI systems on these countries, as well as formulating regulations tailored to their specific context.

As identified in the report of the UN Secretary General's High-level Advisory Body on AI, international governance of AI should be built on the principles of inclusivity and be governed in the public interest and by international human rights law. Among the functions of an international governance structure are conducting research, reinforcing interoperability and developing standards, facilitating AI deployment for societal benefit, collaboration on AI datasets, systems and talent, risk monitoring and emergency response, and establishing norms.[316]

While various international organizations, including the United Nations, UNESCO, G7, OECD, OSCE and the Council of Europe, have initiated efforts to establish global principles and set up an international AI governance structure, these initiatives do not always align or complement each other. To maximize their impact, greater harmonization and more active collaboration among these organizations is needed.

In addition, democratic countries should enhance cooperation using the Partnership for Information and Democracy to establish best practices and global standards and principles regarding AI systems and their impact on the information space. A Global AI Forum inspired by the governance structure of Internet Corporation for Assigned Names and Numbers (ICANN) could serve as a place for inclusive dialogue and a hub on AI issues such as research, in cooperation with the Observatory on Information and Democracy[317], facilitating the harmonization of international initiatives and promoting international

[316] UN Secretary General's High-level Advisory Body on AI (2023*). Interim Report: Governing AI for Humanity.* Available at: https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/ai_advisory_body_interim_report.pdf (Accessed: 7 February 2024).

[317] Forum on Information and Democracy. *International Observatory on Information and Democracy*. Available at: https://informationdemocracy.org/mission/ (Accessed: 7 February 2024).

principles and legislation. Such an initiative should build upon existing efforts such as the Global Partnership on Artificial Intelligence[318] or the OECD Network of Experts on AI[319] while making sure that civil society, media and journalists and other public interest and community representatives have an equal seat at the table.

RECOMMENDATIONS TO **STATES**

> **Ensure that international AI governance is governed by democratic principles in strengthening cooperation through the Partnership for Information and Democracy to promote research, develop best practices, and establish global standards and principles for the development, deployment, and use of AI systems in the information space.**

> **Ensure that AI governance is built upon the values of fairness, equity and non-discrimination.** In doing so, **it is crucial to adequately address the needs of Global Majority countries,** particularly by:

- Redefining policy targets related to the notion of fairness to encompass issues that are specific to the Global Majority (e.g., cast discrimination).
- Taking measures to tackle inequalities in terms of access to knowledge and resources for the development and deployment of AI systems.
- Supporting the establishment of robust regulatory frameworks such as on privacy.

> Build upon existing initiatives, **promote the formation of a Global AI Forum for Open Dialogue with sustained and equal participation of civil society, media and journalists, researchers, and other community and public interest organizations. The purpose of such a forum is to facilitate an open interchange among stakeholders on issues which are central to AI governance.** While over time this forum should organically evolve into a more structured governance structure, such a dynamic setup would ensure that its structure remains adaptable and inclusive over time, effectively responding to the rapid advancement and complex challenges presented by AI technologies. In this regard, States should draw inspiration from organizations like ICANN (Internet Corporation for Assigned Names and Numbers) for effective coordination in managing internet resources, as well as consider the practices of international standard-setting bodies.
Such a Global AI Forum for Open Dialogue could focus on tasks such as:

- **Acting as a central source for AI research and development, offering expertise and guidance on AI-related issues to national entities** involved in AI governance, including legal and judicial systems. It would cooperate closely with the Observatory on Information and Democracy to conduct meta research on AI's impact on the information space.
- **Facilitating the harmonization of global AI initiatives, fostering alignment, and contributing to the development and dissemination of global AI standards.**
- **Developing and promoting principles and best practices regarding legislation and developing standards, thereby strengthening interoperability, such as those on data provenance and curation practices** (see Chapter, section 1.a).

[318] The Global Partnership on Artificial Intelligence (2020). *About - GPAI.* Available at: https://gpai.ai/about/ (Accessed: 7 February 2024).
[319] OECD. *Working Party and Network of Experts on AI*. Available at: https://oecd.ai/en/network-of-experts (Accessed: 7 February 2024).

- ♦ **Collaborating with international bodies** to address AI challenges within the framework of existing conventions, such as the Budapest Convention on Cybercrimes.
- ♦ **Overseeing trends of AI development and regulation, monitoring risks and anticipating future impact, acting as an international watchdog.**
- ♦ **Fostering cooperation to build and share public alternatives to for-profit AI systems, datasets and infrastructure; building talent and conducting AI literacy campaigns** (see Chapter 3, section 1.5 and section 3.2).

> **Recognize and incorporate the significance of promoting ethical AI systems development and deployment through trade. This can be achieved by:**

- ♦ **Incorporating ethical AI clauses in trade agreements**. Trade agreements should include clauses promoting ethical AI development and deployment, guided by UNESCO's *Recommendation on the Ethics of Artificial Intelligence*.[320] Ethical AI clauses could include:
  - Prohibitions on discriminatory AI practices based on sensitive data.
  - Cooperation on capacity building and knowledge sharing regarding ethical AI practices to create a level playing field, with a focus on Global Majority countries. This could involve joint research initiatives and technology transfer agreements.
- ♦ **Collaborating with international standard-setting bodies to develop standards for ethical AI that could be incorporated in trade agreements,** providing a consistent framework across different jurisdictions.

[320] UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence.* Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

ACKNOWLEDGMENTS

The Forum would like to warmly thank the members of this Working Group, especially its co-chairs and the 12 other experts who accepted to join, for their time, insightful perspective and support throughout the whole process.

The Forum would like to thank the representatives of States and public institutions of the Partnership for Information and Democracy that participated in the consultation meeting held on 28 November 2023 for their insights and perspectives on the draft recommendations.

The Forum would also like to warmly thank the almost 40 experts we interviewed, for their time, insightful perspective and support.

- **Ginny Badanes,** *Democracy Forward, Microsoft*
- **Yoshua Bengio,** *Full Professor, Université de Montréal; Founder and Scientific Director, Mila, Québec Artificial Intelligence Institute*
- **Renato Berrino Malaccorto,** *Open Data Charter*
- **Jamie Berryhill,** *Artificial Intelligence Analyst, OECD*
- **Bruno Bioni,** *Founder and Director, Data Privacy Brasil*
- **Lena-Maria Böswald,** *Policy & Advocacy Manager, Das NETTZ - Vernetzungsstelle gegen Hate Speech*
- **Jared Brown,** *Cofounder, Executive Director, Global Shield*
- **Ethan Chumley,** *Senior Cybersecurity Strategist, Microsoft*
- **Styli Charalambous,** *CEO, Daily Maverick*
- **Colin Crowell,** *Senior Advisor, Common Sense Media; Managing Director, Blue Owl Group, and former VP Global Policy, Twitter*
- **Laura Edelson,** *Assistant Professor of Computer Science, Northeastern University*
- **Alessandro Fabris,** *Postdoctoral Researcher, Max Planck Institute for Security and Privacy*
- **Alberto Fernandez Gibaja,** *Head of Programme, Digitalization and Democracy, International IDEA*
- **Sam Gregory,** *Executive Director, WITNESS*
- **Philipp Hacker,** *Professor for Law and Ethics of the Digital Society, European New School of Digital Studies*
- **Sebastian Hallensleben,** *Head of Digitalisation and AI, VDE Association for Electrical, Electronic & Information Technologies*
- **David Evan Harris,** *Chancellor's Public Scholar, UC Berkeley*
- **Clara Helming,** *Senior Policy & Advocacy Manager, Algorithm Watch*
- **Kristina Irion,** *Associate Professor, Institute for Information Law, University of Amsterdam*
- **Martin Kretschmer,** *Professor of Intellectual Property Law and Director of CREATe Centre, University of Glasgow*
- **Amanda Leal,** *Associate, AI Governance, The Future Society*
- **Lucas Lasota,** *Legal Programme Manager, Free Software Foundation Europe*
- **Michael Markovitz,** *Head of GIBS Media Leadership Think Tank, Gordon Institute of Business Science, University of Pretoria*
- **Jacob Mchangama,** *CEO, Justitia; Director of Justitia´s Future of Free Speech Project*
- **Hanna Möllers,** *Deputy General Secretary, German Journalists' Association Deutsche-Journalisten-Verband*
- **Lucy Mwaura,** *Cofounder/Managing Director, LWM CONSULTING*
- **Lê Nguyên Hoang,** *Founder, Science4all; CEO, Calicarpa*

- **Cédric O,** *Member of the French Committee on Generative AI, former Minister of State for Digital Affairs, Co-founding Advisor, Mistral AI*
- **Helena Puig Larrauri,** *Cofounder & Strategy Lead, Build Up*
- **Sannuta Raghu,** *Editor, Scroll*
- **Marietje Schaake,** *International Policy Director, Stanford University Cyber Policy Center; International Policy Fellow at Stanford's Institute for Human-Centered Artificial Intelligence*
- **Anya Schiffrin,** *Director, Technology, Media, and Communications Specialization, School of International and Public Affairs, Columbia University*
- **Emily Tucker,** *Executive Director, Center on Privacy & Technology, Georgetown Law*
- **Richard Wingfield,** *Director of Technology and Human Rights, BSR*
- **Samuel C. Woolley,** *Assistant Professor in the School of Journalism and an Assistant Professor, by courtesy, in the School of Information, University of Texas at Austin*
- **Antonio Zappulla,** *Chief Executive Officer, Thomson Reuters Foundation*
- **Georg Zoeller,** *Director, AI Literacy & Transformation Institute*

The Forum would like to warmly thank the experts that participated in dedicated expert meetings on "Detection Mechanisms, Watermarking and Content Provenance" held on 15 November 2023 and on "Accountability and Liability" held on 12 December for their time, insightful perspective and support. We would also like to thank the Working Group members that participated in these meetings.

- **Jan Beyer,** *Digital Democracy Research Coordinator, Democracy Reporting International*
- **Przemyslaw Biecek,** *Warsaw University of Technology, Global Partnership on Artificial Intelligence's project on Responsible AI for Social Media Governance*
- **Tapabrata Chakraborty,** *Postdoctoral Researcher, University of Oxford and Turing Institute*
- **Andreas Engel,** *Senior Research Fellow at Heidelberg University*
- **David Eyers,** *Department of Computer Science, University of Otago*
- **Graziana Kastl-Riemann,** *lawyer focused on liability aspects related to copyright and privacy infringement by AI-driven software*
- **Susan Leavy,** *School of Information and Communication Studies, University College Dublin, Global Partnership on Artificial Intelligence's project on Responsible AI for Social Media Governance*
- **Virginia Morini,** *National Research Council of Italy, NIRC*
- **Judy Parnall,** *Project Origin, Head of Standards & Industry, BBC*
- **Ugo Pagallo,** *former lawyer and Professor of Jurisprudence, Department of Law, University of Turin; member of the Expert Group set up by the EU Commission on Liability*
- **Beatriz Saab,** *Digital Democracy Researcher, Democracy Reporting International*
- **Paul D. Teal,** *School of Engineering and Computer Science, Victoria University of Wellington*
- **Andrew Trotman,** *Department of Computer Science, University of Otago*

The Forum would like to warmly thank the experts who sent a written contribution for the report with their insightful perspectives.

- **Ayse Arslan,** *Oxford/Cambridge Alumni, Silicon Valley Chapter*
- **Merrin Muhammed Ashraf,** *Research Assistant, IT for Change, Bengaluru*
- **Paul M. Barrett,** *Deputy Director and Senior Research Scholar NYU Stern Center for Business and Human Rights*
- **Yoshua Bengio,** *Full Professor, Université de Montréal, Founder and Scientific Director, Mila, Québec Artificial Intelligence Institute*
- **Jennifer Brody,** *Deputy Director, Policy and Advocacy Technology and Democracy, Freedom House*
- **Luo Chen Jun,** *Innovator and Advocate, online safety and content moderation*
- **Laurence Dierickx,** *Postdoctoral Researcher, University of Bergen*
- **José van Dijck,** *Professor of Media and Digital Society, Utrecht University*
- **Michael P. Goodyear,** *Acting Assistant Professor, New York University School of Law*
- **Arijit Goswami,** *Senior Innovation Manager at a large technology services institute*
- **Philipp Hacker,** *Professor for Law and Ethics of the Digital Society, European New School of Digital Studies*
- **Mike Harris,** *Exonym*
- **Daniel Innerarity,** *Chair in AI & Democracy, Florence School of Transnational Governance, European University Institute*
- **Jordan Ogg,** *International Policy Manager, Ofcom*
- **Hassan Patel,** *Director of Global AI Policy Compliance Engineering, Fairly AI*
- **Maria Paz Canales,** *Jacqueline Rowe, Ian Barber, Global Partners Digital*
- **Ioannis Pitas,** *Professor, Department of Informatics, University of Thessaloniki (AUTH); Director of the Artificial Intelligence and Information Analysis (AIIA) lab*
- **Nicholas Sewe,** *Engagement Manager, Digital Inclusion Benchmark, World Benchmarking Alliance (WBA)*
- **Nii Simmonds,** *Senior Fellow at Atlantic Council's GeoTech center*
- **James Wan,** *Berserq*

To strengthen the participation of experts from different regions in the Working Group, the Forum formalized a partnership with eight local organizations in Western Africa, South Africa, Lebanon and Latin America:

- **Association for the Fight against Racism, Ethnocentrism and Regionalism** *(ALCRER), a non-governmental organization defending human rights, democracy and good governance in Benin*
- **Community Focus Foundation Ghana is a civil society organization working on governance,** *participation and policy reform*
- **Data Leads,** *an Indian digital media and tech company*
- **Jonction,** *a Senegalese non-governmental organization working on human rights, data protection, cybersecurity and online speech*
- **Maharat Foundation,** *a non-governmental organization for freedom of expression and media development based in Lebanon*
- **Media Monitoring Africa,** *is a South African organization promoting ethical and fair journalism*
- **Network of Online Press Professionals of Côte d'Ivoire (REPPRELCI),** *the first professional organization of the digital press in Ivory Coast*
- **Observacom (Latin American Observatory of Regulation,** *Media and Convergence) a non-profit, professional and independent regional think tank, specialized in regulation and public policies related to the media, telecommunications, internet and freedom of expression*

Thanks to their support, 61 experts were mobilized and participated in workshops to develop the recommendations presented in this report:

Benin

- **Marius Janvier Dossou-Yovo,** *Civil Administrator, Doctor of State in Private Law, Expert in Digital and Media Law, Coordinator of the think tank Information Society: Practices and Governance, Professor and Researcher*
- **Wenceslas Mahoussi,** *Senior Lecturer Information and Communication Sciences, Ecole Nationale des Sciences et Techniques de l'Information et de la Communication (ENSTIC), University of Abomey-Calavi*
- **Yisségnon Rémy Oke,** *Information Technology Infrastructure Engineer, Director of Digital, Ministry of Digital and Digitalization, and National Technical Assistant for the Digital Sector, Enabel - Belgium Development Agency*
- **Toundé Seth Amon Dedehouanou,** *Information Technology Infrastructure Engineer, Ministry of Digital Affairs and Digitalization*
- **Alain Codjo Cakpo,** *Information Technology Infrastructure Engineer, Head of Information Technology Service, General Directorate of Higher Education*
- **Vinasétan Ratheil Esse Houndji,** *Senior Lecturer in Artificial Intelligence, University of Abomey-Calavi (UAC); Head of the Software Engineering Department, Institute for Training and Research in Computer Science, UAC; Chair, Ratheil Foundation for Responsible and Efficient Artificial Intelligence*
- **Gérard Nakou,** *Information Technology Infrastructure Engineer, Information Systems Management, Pehunco and Kouande local government*
- **Hans Norbert Atacle,** *Project Manager, ONG ALCRER, Operation Research, Institut de Mathématiques et de Sciences Physiques, Cotonou*
- **Gervais Loko,** *Expert in Governance and Democracy and Program Manager, ONG ALCRER*

Ghana

- **Richard Kasu,** *Policy Analyst, CFF-Ghana*
- **Martin Thompson Ntem,** *Lecturer and Digital Communications Specialist, Institute for Digital Marketing Communications, Ghana*
- **Miriam Ocloo,** *Tech Trainer, EM Services, Ghana*
- **Francis Kasu,** *Population Health Informatics Specialist, CFF-Ghana*
- **Abraham Dzagbletey,** *Lecturer and Digital Marketer, Institute for Digital Marketing Communications, Ghana*
- **Princess Lovia Tetteh,** *Internet Governance and Youth Development Specialist, Love Aid Foundation*
- **Albert Gharbin,** *Communication Design Specialist, McGharbins Group*
- **Charles Obianim,** *IT & Financial Specialist, CFF-Ghana*

India

- **Prof. Saima Saeed,** *Associate Professor, Centre for Media, Law and Governance, Jamia Millia Islamia (JMI), New Delhi*
- **Prof. Taberez Neyazi,** *Assistant Professor of Political Communication and New Media & Director of Digital Campaign Asia Project, National University of Singapore*
- **S.Y. Qureshi,** *Former Chief Election Commissioner, India*
- **Arpit Chaturvedi,** *Cofounder and CEO, Global Policy Insights*
- **Manisha Pathak Shelat,** *Professor, Communication & Digital Platforms and Strategies, MICA, Ahmedabad, India*
- **Sam Daniels,** *Senior Television Journalist, AI Enthusiast*

- **Nisha Bhambani,** *Senior Advocate, Supreme Court, India*
- **Kavya Sukumar,** *Principal (Civic Technology and Media Sectors), Lightrock*

Ivory Coast

- **Lassina Serme,** *Senior Journalist, President, Réseau des Professionnels de la Presse en Ligne de Côte d'Ivoire (REPPRELCI), Expert on Digital Issues*
- **Karim Wally,** *Journalist; Lecturer, Université Félix Houphouët-Boigny*
- **Dr Achi Harrison,** *Director, Métaverse Laboratory, Université Virtuelle de Côte d'Ivoire*
- **Lucien Houedanou,** *President, Cénacle des Journalistes Seniors de Côte d'Ivoire*
- **Mamadou Konate,** *Data Scientist, Developer*

Latin America

- **Flavia Costa,** *Adjunct Researcher, National Council for Scientific and Technical Research (CONICET), Argentina*
- **Martín Becerra,** *Professor, Universidad Nacional de Quilmes (UNQ); Chief Researcher, National Council for Scientific and Technical Research (CONICET), Argentina*
- **Patricia Peña,** *Director, Datos Protegidos, Chile*
- **Juan Ortiz,** *Researcher, Berkman Klein Center for Internet & Society, Harvard University, and Founder of Common Ground, Argentina and United States*
- **Luis Fernando García,** *Executive Director, Red en Defensa De Los Derechos Digitales (R3D), Mexico*
- **Edison Lanza,** *former Special Rapporteur for Freedom of Expression, Inter-American Commission of Human Rights, Uruguay*
- **Paulina Gutiérrez,** *former Digital Rights Program Officer at ARTICLE19, Mexico*
- **Maia Levy Daniel,** *Research Affiliate, Center of Technology and Society (CETyS), Universidad de San Andrés, Argentina, and former Director of Research and Public Policy, Centro Latam Digital, Mexico*
- **Ramiro Álvarez Ugarte,** *Vice-Director of the Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE), Argentina*

Lebanon

- **Dr Maria Bou Zeid,** *Dean of the Faculty of Humanities, Notre Dame University- Louaize (NDU)*
- **Layal Jebran,** *Technology Expert*
- **Zeina Bou Harb,** *International Cooperation Manager, OGERO Telecom; Head of the General Secretariat, Lebanese Internet Governance Forum*
- **Abed Kataya,** *Digital Content Manager, SMEX*
- **Tony Mikhael,** *Legal Expert*
- **Dr Marc Ibrahim,** *Director, Institut National des Télécommunications et de l'Informatique (INCI); Associate Professor, Ecole Supérieure des Ingénieurs de Beyrouth (ESIB), Saint-Joseph University of Beirut*
- **Wael Akiki,** *Program Manager, Samir Kassir Foundation*
- **Layal Sakr,** *Legal Expert, Founder of Seeds for Legal Initiatives*

Senegal

- **Ndeye Fatou Mboup,** *Junior Expert Researcher, IPAR/World Partnership on Artificial Intelligence*
- **Justin Oumar Bamahossovi,** *Lawyer in charge of cooperation, Commission for the Protection of Personal Data CDP; Researcher in international cyberspace law*
- **Fana Cissé,** *Journalist/Reporter, Media/PressAfrik*
- **Maateuw Mbaye,** *Programme Officer, Article 19*

- **Mouhamed Ndiaye Bocoum,** *Legal Consultant, Commission for the Protection of Personal Data (CDP)*
- **Abdoulaye Diallo,** *Coordinator (Legal and Scientific Information), Digital Rights Department, Réseau des Professionnels de la Presse en Ligne de Côte d'Ivoire (RADDHO)*
- **Assane Sy,** *Consultant & Trainer, UnLine Sas*
- **Emmanuel Diokh,** *Techno-Pedagogue, Legal Trainer & President, Internet Sans Frontières Senegal*

South Africa

- **Dimitri Martinis,** *CEO, MCM Digital Media*
- **Unathi Malunga,** *Entertainment Lawyer, Creative Industries Consultant and Content Executive*
- **Izak Minnaar,** *Journalism, Digital Media, Elections and Policy Expert, Consultant & Trainer, SA National Editors' Forum (Sanef), SA Press Council, Support Public Broadcasting Coalition (SOS)*
- **Kgothatso Mampa,** *Media Law and Commercial Copyright Expert*
- **Tharin Pillay,** *Fellow, ALT Advisory*
- **Sarah Chiumbu,** *Associate Professor, Department of Communication and Media, University of Johannesburg*

The Forum would like to warmly thank the partner organizations that supported the launch event of this policy framework on 28 February 2024, namely:

- **Center for Human-Compatible Artificial Intelligence** *based at UC Berkeley (United States)*
- **Florence School of Transnational Governance at the European University Institute** *(Italy)*
- **Center for Law, Internet and Society of the Institute for Development, Education and Research** *(Brazil)*
- **Research ICT Africa** *(South Africa)*
- **SciencesPo Paris School of International Affairs, Tech and Global Affairs Innovation Lab** *(France)*

BIBLIOGRAPHY

Allen, D. and Weyl, E.G. (2024). The Real Dangers of Generative AI. *Journal of Democracy*. Available at: www.journalofdemocracy.org/articles/the-real-dangers-of-generative-ai/ (Accessed: 7 February 2024).

Amazon Web Services (2023). *What is an API? - API Beginner's Guide - AWS*. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/what-is/api/. (Accessed: 2 Feb 2024)

Amnesty International (2022). *Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – Report.* Available at: www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ (Accessed: 7 February 2024).

Anderljung, M., Smith, E., O'brien, J., Soder, L., Bucknall, B., Bluemke, E., Schuett, J., Trager, R., Strahm, L. and Chowdhury, R. (2023). *Towards Publicly Accountable Frontier LLMs.* Available at: https://arxiv.org/pdf/2311.14711.pdf (Accessed: February 7 2024).

Arguedas, A., Robertson, C., Fletcher, R. and Nielsen, R. (2022). *Echo chambers, filter bubbles, and polarisation: a literature review.* Reuters Institute for the Study of Journalism. Available at: https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review (Accessed: 7 February 2024).

Arisoy, E., Leipold, K. and Messan, K. (2023), The expanding role of public procurement in Africa's economic development, World Bank Blogs, available at: https://blogs.worldbank.org/governance/expanding-role-public-procurement-africas-economic-development (Accessed on 9 February 2024).

Awan, A.A. (2022). *A Complete Guide to Data Augmentation*. [online] www.datacamp.com. Available at: www.datacamp.com/tutorial/complete-guide-data-augmentation (Accessed: 31 January 2024).

ASEAN (2023). ASEAN *Guide on AI Governance and Ethics.* Available at: https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf (Accessed: 7 February 2024).

Associated Press (2024). *AI-powered misinformation is the world's biggest short-term threat, Davos report says.* Available at: https://apnews.com/article/artificial-intelligence-davos-misinformation-disinformation-climate-change-106a1347ca9f987bf71da1f86a141968 (Accessed: 7 February 2024).

Bell, A., Stoyanovich, J. and Nov, O. (2023). *Algorithmic Transparency Playbook*. Center for Responsible AI. Available at: https://dataresponsibly.github.io/algorithmic-transparency-playbook/resources/transparency_playbook_camera_ready.pdf (Accessed: February 7 2024).

Belli, L. et al. (2022), Towards meaningful and interoperable transparency for digital platforms, Internet Governance Forum, available at: www.intgovforum.org/en/filedepot_download/57/23886 (Accessed: 7 February 2024).

Bengani, P., Stray, J., & Thorburn, L. (2022). *Blog Post: What's Right and What's Wrong with Optimizing for Engagement.* Center for Human-Compatible AI at UC Berkeley. Available at: https://humancompatible.ai/news/2022/05/02/blog-post-whats-right-and-whats-wrong-with-optimizing-for-engagement/ (Accessed: 7 February 2023).

Bengio, Y. (2023). *AI and Catastrophic Risk*. Journal of Democracy. Available at: www.journalofdemocracy.org/ai-and-catastrophic-risk/ (Accessed: February 7 2024).

Bertuzzi, L. (2024). *EU countries give crucial nod to first-of-a-kind Artificial Intelligence law.* Euractiv. Available at: www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/ (Accessed: 7 February 2024).

Birhane, A. et al (2021). *Multimodal datasets: misogyny, pornography, and malignant stereotypes.* Available at: https://arxiv.org/abs/2110.01963 (Accessed: 7 February 2024).

Bitdefender, *What is social media impersonation?* Available at: www.bitdefender.com/cyberpedia/what-is-social-media-impersonation/ (Accessed on 8 February 2024).

Bizga, A. (2020), *Blog Post: What is impersonation?,* Bitedefender. Available at: www.bitdefender.com/blog/hotforsecurity/what-is-impersonation/ (Accessed on 8 February 2024).

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D. and Liang, P. (2023). *The Foundation Model Transparency Index*. Available at: https://crfm.stanford.edu/fmti/fmti.pdf (Accessed: February 7 2024).

Bontcheva, K. et al (2020), *Legislative and Regulatory Responses to Disinformation, Excerpt from the Original Report*, Broadband Commission for Sustainable Development. Available at: https://en.unesco.org/sites/default/files/balanceact_legislative_en.pdf (Accessed: 8 February 2024).

Brookings. (n.d.). *Detecting AI fingerprints: A guide to watermarking and beyond*. [online] Available at: www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/ (Accessed: 31 January 2024).

BSA (2023). *AI Developers and Deployers: An Important Distinction | BSA | The Software Alliance*. [online] Available at: www.bsa.org/policy-filings/ai-developers-and-deployers-an-important-distinction (Accessed: 2 February. 2024).

Cambridge Consultants (2019). *Use of AI in Online Content Moderation.* Available at: www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Accessed: 7 February 2024).

Chang, T.A., Arnett, C., Tu, Z. and Bergen, B.K., (2023). *When is multilinguality a curse? Language modeling for 250 high-and low-resource languages.* Available at: https://doi.org/10.48550/arXiv.2311.09205 (Accessed: 7 February 2024).

Coleman, J.L. (2002). Fault and strict liability. *Risks and Wrongs*, pp.212–233. Doi: https://doi.org/10.1093/acprof:oso/9780199253616.003.0012 (Accessed: 8 February 2024).

Content Authenticity Initiative (n.d.). *How it works.* Available at: https://contentauthenticity.org/how-it-works (Accessed: 7 February 2024).

Council of Europe (2023). *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law.* Available at: https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043 (Accessed: 7 February 2024).

Council of Europe (2023). *Guidelines on the responsible implementation of artificial intelligence systems in journalism*. Available at: https://rm.coe.int/cdmsi-2023-014-guidelines-on-the-responsible-implementation-of-artific/1680adb4c6 (Accessed: 7 February 2024).

c2pa.org. (n.d.). *FAQ - C2PA*. [online] Available at: https://c2pa.org/faq/ (Accessed: 31 January 2024).

DataLeads. *FactShala: India's Largest Media Literacy Network*. available at: https://dataleads.co.in/capacity-building/#FactShala (Accessed on 9 February 2024).

Delcker, J. (2019), *Finland's grand AI experiment*, Politico, available at: www.politico.eu/article/finland-one-percent-ai-artificial-intelligence-courses-learning-training/ (Accessed on 9 February 2024).

DeepAI. (2019). *Classifier*. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/classifier (Accessed: 2 Feb 2024).

Díaz, A. and Hecht-Felella, L. (2021). *Double Standards in Social Media Content Moderation,* Brennan Center for Justice. Available at: www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf (Accessed: 7 February 2024).

Edelson, L., Haugen, F., and McCoy, D. (2023). *Demystifying Social Media Feeds: A Taxonomy and Transparency for Algorithmic Feed Systems Designs.* Draft manuscript.

Eisenstein, J. (2019). *Introduction to natural language processing*. Cambridge, Massachusetts: The MIT Press.

European Commission (2024). *Commission is gathering views on draft DSA guidelines for election integrity*. Available at https://digital-strategy.ec.europa.eu/en/news/commission-gathering-views-draft-dsa-guidelines-election-integrity (Accessed: 15 February 2024).

European Commission (2024). *Commission Decision Establishing the European AI Office.* Available at: https://digital-strategy.ec.europa.eu/en/library/commission-decision-establishing-european-ai-office (Accessed: February 7 2024).

European Commission (2023). *Delegated Regulation on independent audits under the Digital Services Act.* Available at: https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act (Accessed: February 7 2024).

European Commission (2023). *Hiroshima Process International Code of Conduct for Advanced AI Systems.* Available at: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems (Accessed: 7 February 2024).

European Commission (2022). *AI Watch: Estimating AI Investments in the European Union*. Available at: https://ai-watch.ec.europa.eu/publications/ai-watch-estimating-ai-investments-european-union_en (Accessed: February 7 2024).

European Commission (2022). *European Centre for Algorithmic Transparency*. Available at: https://algorithmic-transparency.ec.europa.eu/about_en (Accessed: February 7 2024).

European Commission (2019). Ethics guidelines for trustworthy AI. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed: 15 February 2024).

European Commission (2018), *A Definition of AI: Main Capabilities and Scientific Disciplines*, p.1. Available at: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (Accessed: 8 February 2024)

European Council (2024). *Artificial Intelligence Act: Council and Parliament strike a deal on the first rules for AI in the world.* Available at: www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ (Accessed: 7 February 2024).

EUR Lex (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).* Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065 (Accessed: 7 February 2024).

European Parliament (2022). *Auditing the quality of datasets used in algorithmic decision-making systems.* Available at: www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf (Accessed: 7 February 2024).

European Parliament (2020), *European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence*, available at: www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.html (Accessed on 8 February 2024).

European Union (2019), *Liability for artificial intelligence and other emerging digital technologies,* available at: https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en (Accessed: 7 February 2024).

faculty.washington.edu. (n.d.). *About Data Provenance*. [online] Available at: https://faculty.washington.edu/hazeline/ProvEco/generic.html (Accessed: 1 February 2024).

Fair Trade International. About Us. Available at: www.fairtrade.net/about (Accessed: 7 February 2024).

Forum on Information and Democracy (2024), *Protecting Democratic Elections through Safeguarding Information Integrity*. Forum on Information & Democracy, International IDEA, Democracy Reporting International. Available at: https://informationdemocracy.org/2024/01/30/tech-firms-governments-urged-to-combat-digital-election-threats/ (Accessed: 7 February 2024).

Forum on Information and Democracy (2023). *Pluralism on News and Information in Curation and Indexing Algorithms*. Available at: https://informationdemocracy.org/wp-content/uploads/2023/02/Report_Pluralism-in-algorithms.pdf (Accessed: 8 February 2024).

Forum on Information and Democracy (2022). *Accountability Regimes for Social Networks and their Users*, available at: https://informationdemocracy.org/wp-content/uploads/2023/01/ID_Report-on-Accountability-regime_Sept22.pdf (Accessed: 8 February 2024).

Forum on Information and Democracy (2020). *How to End Infodemics*. Available at: https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf (Accessed: 8 February 2024).

Forum on Information and Democracy. *International Partnership for Information & Democracy*. Available at: https://informationdemocracy.org/international-partnership-on-information-democracy/ (Accessed: 8 February 2024).

Forum on Information and Democracy (n.d.) International Observatory on Information and Democracy. Available at: https://informationdemocracy.org/mission/ (Accessed: 7 February 2024).

Freckleton, I. (2020), *COVID-19: Fear, quackery, false representations and the law*, International Journal of Law and Psychiatry, Volume 72. Available at: https://doi.org/10.1016/j.ijlp.2020.101611 (Accessed on 8 February 2024).

FSFE - Free Software Foundation Europe (n.d.). *What is Free Software*. Available at: https://fsfe.org/freesoftware/freesoftware.en.html (Accessed: 7 February 2024).

Frontier Model Forum (2023). *What is Red Teaming?* Available at: www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf (Accessed: 7 February 2024).

Fung, D.O., Brian (2023), *First on CNN: Biden campaign prepares legal fight against election deepfakes*, CNN Politics. Available at: https://edition.cnn.com/2023/11/30/politics/biden-campaign-prepares-against-deepfakes/index.html (Accessed: 8 February 2024).

Funke, D. and Flamini, D. (2018), *A guide to anti-misinformation actions around the world*, Poynter. Available at: www.poynter.org/ifcn/anti-misinformation-actions/ (Accessed: 8 February 2024).

Future of Life (2017). Asilomar AI Principles. Available at: https://futureoflife.org/open-letter/ai-principles/ (Accessed: 7 February 2024).

Gallo, A. (2017). *A Refresher on A/B Testing*. Harvard Business Review. Available at: https://hbr.org/2017/06/a-refresher-on-ab-testing (Accessed: February 7 2024).

GeeksforGeeks, *Robustness Testing*. Available at: www.geeksforgeeks.org/robustness-testing (Accessed: 9 February 2024).

Global Infodemic Management Course. Global Infodemic Management Course for Healthcare Workers. available at: https://gimch.org/ (Accessed on 9 February 2024).

Goldstein, J., Sastry, G., Musser, M., Diresta, R., Gentzel, M., Sedova, K., Adler, S., Avin, S., Bansemer, J., Bregler, C., Brundage, M., Gregory, S., Grossman, S., Herbert-Voss, A., Jernite, Y., Leibowicz, C., Leahy, C., Lin, H., Lohn, D. and Mitchell, M. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. Available at: https://arxiv.org/pdf/2301.04246.pdf (Accessed: 7 February 2024).

Gordon Institute of Business and Science (2023). *Big Tech and Journalism - Principles for Fair Compensation.* Available at: www.gibs.co.za/news-events/news/pages/big-tech-and-journalism-principles.aspx (Accessed: 7 February 2024).

Government of Canada (2023). *Artificial Intelligence and Data Act.* Available at: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act (Accessed: 7 February 2024).

GPAI (2023). *Crowdsourcing the curation of the training set for harmful content classifiers used in social media A pilot study on political hate speech in India, Report,* Global Partnership on AI. Available at: https://gpai.ai/projects/responsible-ai/RAI04%20-%20Crowdsourcing%20the%20Curation%20of%20the%20Training%20Set%20for%20Harmful%20Content%20Classifiers%20Used%20in%20Social%20Media.pdf (Accessed: 7 February 2024)

GPAI (2022). *Transparency Mechanisms for Social Media Recommender Algorithms: from Proposals to Actions. Report,* Global Partnership on AI. Available at: https://gpai.ai/projects/responsible-ai/transparency-mechanisms-for-social-media-recommender-algorithms.pdf (Accessed: 8 February 2024).

Grynbaum, M.M. and Mac, R. (2023). *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*. The New York Times. Available at: www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (Accessed: 7 February 2024).

Halm, K.C.,Kumar, A., Segal, J. and Kalinowski, C. (2020).*Two New California Laws Tackle Deepfake Videos in Politics and Porn. Available at:* www.dwt.com/blogs/media-law-monitor/2020/02/two-new-california-laws-tackle-deepfake-videos-in (Accessed: 12 February 2024).
Henderson, P. (2023), *Who Is Liable When Generative AI Says Something Harmful?* Stanford University, available at:

https://hai.stanford.edu/news/who-liable-when-generative-ai-says-something-harmful (Accessed: 8 February 2024).

Hsu, T., Thompson, S.A. and Myers, S.L. (2024). *Elections and Disinformation Are Colliding Like Never Before in 2024.* The New York Times. Available at: www.nytimes.com/2024/01/09/business/media/election-disinformation-2024.html (Accessed: 7 February 2024).

IBM (n.d.). *What is an AI model? | IBM*. [online] Available at: www.ibm.com/topics/ai-model. (Accessed: 2 Feb 2024)

ICPSR (n.d.). *Data Management & Curation*. [online] Available at: www.icpsr.umich.edu/web/pages/datamanagement/index.html (Accessed: 8 February 2024).

Information Commissioner's Office. *A Guide to ICO Audit Artificial Intelligence (AI) Audits Contents*. Available at: https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf (Accessed: February 7 2024).

intellabs.github.io. (n.d.). *Knowledge Distillation - Neural Network Distiller.* Available at: https://intellabs.github.io/distiller/knowledge_distillation.html (Accessed: 7 February 2024).

ISACA (2018). *Auditing Artificial Intelligence*. Available at: https://ec.europa.eu/futurium/en/system/files/ged/auditing-artificial-intelligence.pdf (Accessed: February 7 2024).

Jernite, Y; (2023). *Training Data Transparency in AI: Tools, Trends, and Policy Recommendations, Hugging Face Community Blog.* Available at: https://huggingface.co/blog/yjernite/data-transparency#data-transparency-in-focus-what-is-needed (Accessed: 7 February 2024).

Ji, J., et al. (2023). *AI Alignment: A Comprehensive Survey*. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2310.19852. (Accessed 2 February 2024)

Journalism Trust Initiative. Available at: www.journalismtrustinitiative.org/ (Accessed: 7 February 2024).

Karathanasis, A.L., Stephanie Celis J. and Theodoros (2022). *Civil Liability for AI Systems: Comment on EU Commission's Proposals.* MIAI. Available at: https://ai-regulation.com/eu-commission-proposals-on-ai-civil-liability/ (Accessed: 7 February 2024).

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. and Goldstein, T. (n.d.), *A Watermark for Large Language Models.* Available at: https://arxiv.org/pdf/2301.10226.pdf (Accessed: 7 February 2024).

Kovic, M., Rauchfleisch, A., Sele, M. and Caspar, C. (2018). *Digital astroturfing in politics: Definition, typology, and countermeasures. Studies in Communication Sciences*, 18(1). Available at: www.hope.uzh.ch/scoms/article/view/j.scoms.2018.01.005/991 (Accessed: 7 February 2024).

Kreps, S. and Kriner, D. (2023). *How AI Threatens Democracy. Journal of Democracy*. Available at: www.journalofdemocracy.org/articles/how-ai-threatens-democracy/ (Accessed: 7 February 2024).

Kretschmer, M., Kretschmer, T., Peukert, A. and Peukert, C. (2023). *The risks of risk-based AI regulation: taking liability seriously. Social Science Research Network*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4622405 (Accessed: February 7 2024).

Kulick, A. (2022), *Meta's Oversight Board and Beyond – Corporations as Interpreters and Adjudicators of International Human Rights Norms,* The Law and Practice of International Courts and Tribunals 2022, Forthcoming, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4226521 (Accessed on 9 February 2024).

Laux, J. (2023). *Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act*. *SSRN Electronic Journal*. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4377481 (Accessed: February 7 2024).
Law Insider. (n.d.). *deployer Definition*. [online] Available at: www.lawinsider.com/dictionary/deployer (Accessed: 2 Feb. 2024).

Leibowicz, C. (2023). *Why watermarking AI-generated content won't guarantee trust online*. MIT Technology Review. Available at: www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/ (Accessed: 7 February 2024).

Lyons, K. (2020). *Facebook rolls back 'nicer' News Feed that boosted mainstream publishers*, The Verge. Available at: www.theverge.com/2020/12/17/22180259/facebook-news-feed-change-post-election-publishers-misinformation (Accessed: 7 February 2024).

Magrani, E. (2019). *New perspectives on ethics and the laws of artificial intelligence.* Internet Policy Review. Available at: https://policyreview.info/articles/analysis/new-perspectives-ethics-and-laws-artificial-intelligence (Accessed: February 7 2024).

Malgieri, G. and Pasquale, F. (2024). *Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. Computer Law & Security Review*. Doi: https://doi.org/10.1016/j.clsr.2023.105899.

Martineau, K. (2021). *What is synthetic data?* [online] IBM Research Blog. Available at: https://research.ibm.com/blog/what-is-synthetic-data (Accessed: 31 January 2024).

Marwala, T. (2023). *Algorithm Bias — Synthetic Data Should Be Option of Last Resort When Training AI Systems.* United Nations University. Available at: https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems (Accessed: 7 February 2024).

Marzotto, M. (2023), *Whistleblowers are human rights defenders. So why don't we protect them like they are?* The Signals Network, available at:
https://thesignalsnetwork.org/whistleblowers-are-human-rights-defenders/ (Accessed on 9 February 2024).

Mehta, R. (2023). *Human Data Labeling for Successful AI*. [online] iMerit. Available at: https://imerit.net/blog/human-data-labeling-for-successful-ai/ (Accessed: 9 February 2024).

Metaxa, D., Park, J.S., Robertson, R.E., Karahalios, K., Wilson, C., Hancock, J. and Sandvig, C. (2021). *Auditing Algorithms: Understanding Algorithmic Systems from the Outside In.* Available at: https://hci.stanford.edu/publications/2021/FnT_AuditingAlgorithms.pdf (Accessed: February 7 2024).

Miller, K and Lohn, A. (2023). *Techniques to Make Large Language Models Smaller: AN Explainers, Center for Security and Emerging Technologies.* Available at: https://cset.georgetown.edu/publication/techniques-to-make-large-language-models-smaller-an-explainer/ (Accessed: 7 February 2024).

Miller, K. (2023), Introducing The Foundation Model Transparency Index, Stanford University, available at:
https://hai.stanford.edu/news/introducing-foundation-model-transparency-index (Accessed: 9 February 2024).

Miller, T., Baird, T., Littlefield, C., Kofinas, G., Chapin, F. and Redman, C. (2008). *Epistemological Pluralism: Reorganizing Interdisciplinary Epistemological Pluralism: Reorganizing Interdisciplinary Research Research. Ecology and Society*. Available at: https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1034&context=unf_research (Accessed: February 7 2024).

Milmo, D. (2023), *AI firms 'should include members of public on boards to protect society', The Guardian*, available at:
www.theguardian.com/technology/2023/dec/06/ai-firms-should-include-members-of-public-on-boards-to-protect-society (Accessed: 9 February 2024).

Mislove, A. (2023). *OSTP Blog Post: Red-Teaming Large Language Models to Identify Novel AI Risks.* The White House, Office of Science and Technology Policy. Available at: www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/ (Accessed: 7 February 2024).

Mökander, J., Axente, M., Casolari, F. and Floridi, L. (2021). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, 32. Doi: https://doi.org/10.1007/s11023-021-09577-4.

Mulder, V., Mermoud, A., Lenders, V., Tellenbach, B. (eds) *Trends in Data Protection and Encryption Technologies*. Springer, Cham. Available at: https://doi.org/10.1007/978-3-031-33386-6_6 (Accessed: 7 February 2024).

Nanni, D. (2023). *Synthetic Entities: Definitions, Characteristics, and Future Perspectives*. [online] Brass For Brain. Available at: https://medium.com/brass-for-brain/synthetic-entities-definitions-characteristics-and-future-perspectives-49673f22f6fe (Accessed: 31 January 2024).

News/Media Alliance (2023). *White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement And Not a Fair Use.* Available at: www.newsmediaalliance.org/generative-ai-white-paper/ (Accessed: 7 February 2024).

Nicholas, G. and Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis*. Center for Democracy and Technology. Available at: https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ (Accessed: 7 February 2024).

NIST (n.d.). *Computer Security Resource Center - Glossary*. [online] csrc.nist.gov. Available at: https://csrc.nist.gov/glossary/term/Red_Team (Accessed: 8 February 2024).

OECD. *OECD AI Principles overview*. Available at: https://oecd.ai/en/ai-principles (Accessed: February 7 2024).

OECD. *Algorithmic Transparency Recording Standard*. Observatory of Public Sector Innovation. Available at: https://oecd-opsi.org/innovations/algorithmic-transparency-standard/ (Accessed: February 7 2024).

OECD. *OECD Working Party and Network of Experts on AI*. Available at: https://oecd.ai/en/network-of-experts (Accessed: February 7 2024).

OECD (2024), *Size of Public Procurement, Government at a Glance 2023*, available at www.oecd-ilibrary.org/sites/3d5c5d31-en/1/3/7/1/index.html?itemId=/content/publication/3d5c5d31-en&_csp_ (Accessed: 9 February 2024).

OECD (2023). *Official Development Assistance (ODA) - OECD*. Available at: www.oecd.org/dac/financing-sustainable-development/development-finance-standards/official-development-assistance.htm (Accessed: February 7 2024).

OECD (2023). *Outcome Statement on the Two-Pillar Solution to Address the Tax Challenges Arising from the Digitalisation of the Economy*, available at www.oecd.org/tax/beps/outcome-statement-on-the-two-pillar-solution-to-address-the-tax-challenges-arising-from-the-digitalisation-of-the-economy-july-2023.pdf (Accessed on 9 February 2024).

OECD (2022). *Responsible AI licenses: a practical tool for implementing the OECD Principles for Trustworthy AI*. Available at: https://oecd.ai/en/wonk/rails-licenses-trustworthy-ai (Accessed: February 7 2024).

OECD (2020), *Government at a Glance: Latin America and the Caribbean 2020*, available at: www.oecd.org/publications/government-at-a-glance-latin-america-and-the-caribbean-5ceda53e-en.htm (Accessed: 9 February 2024).

OECD (2019). *Recommendation of the Council on Artificial Intelligence.* Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed: 7 February 2024).

OECD (2019). *The OECD Artificial Intelligence (AI) Principles.* Available at: https://oecd.ai/en/ai-principles. (Accessed: 7 February 2024).

OECD (2015), *Recommendation of the Council on Public Procurement*, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0411 (Accessed: 9 February 2024).

Ogunleye, I. (2022). *AI's Redress Problem: Recommendations to Improve Consumer Protection from Artificial Intelligence*. Available at: https://cltc.berkeley.edu/publication/cltc-white-paper-ais-redress-problem/ (Accessed: February 7 2024).

OHCHR (n.d.) Access to remedy and the technology sector: basic concepts and principles, UN B-Tech Foundational Paper available at: www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf (Accessed: 8 February 2024).

OHCHR (n.d.), Taxonomy of Human Rights Risks Connected to Generative AI, available at: www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf (Accessed: 7 February 2024).

OpenAI (2023), *Privacy policy.* Available at: https://openai.com/policies/privacy-policy (Accessed: February 7 2024).

Oracle.com (2022). *What is A/B Testing?* Available at: www.oracle.com/cx/marketing/what-is-ab-testing/ (Accessed 2 February 2024)

OSCE (2021), *Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual*, available at: www.osce.org/files/f/documents/8/f/510332_1.pdf (Accessed: 7 February 2024)

Oversight Board (n.d.). Available at: www.oversightboard.com/ (Accessed: 7 February 2024).

Partnership on AI (n.d.) PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action, available at: https://syntheticmedia.partnershiponai.org/ (Accessed: 7 February 2024).

Project Origin (n.d.). *Project Origin.* Available at: www.originproject.info/about (Accessed: 7 February 2024).

Public Citizen (2023), *Comment to FEC: A.I.-Generated Political Deepfakes Are 'Fraudulent Misrepresentation.'* Available at: www.citizen.org/article/comment-to-fec-a-i-generated-political-deepfakes-are-fraudulent-misrepresentation/ (Accessed: 7 February 2024).

Responsible AI Licenses (n.d.) available at www.licenses.ai/ (Accessed: 7 February 2024)

Robertson, D. (2023). *When 'red-teaming' AI isn't enough.* Politico. Available at: www.politico.com/newsletters/digital-future-daily/2023/10/25/when-red-teaming-ai-isnt-enough-00123577 (Accessed: 7 February 2024).

Routley, N. (2023). *What is generative AI? An AI explains*. [online] World Economic Forum. Available at: www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/ (Accessed: 9 February 2024).

RSF (2023), Paris Charter on AI and Journalism, available at: https://rsf.org/en/rsf-and-16-partners-unveil-paris-charter-ai-and-journalism (Accessed: 7 February 2024).

RSF (2023), *RSF and French general press alliance launch "Spinoza Project" to develop AI tool for journalists*. Available at: https://rsf.org/en/rsf-and-french-general-press-alliance-launch-spinoza-project-develop-ai-tool-journalists (Accessed: 9 February 2024).

Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N. and Allen, P. (2019). *The Risk of Racial Bias in Hate Speech Detection.* Association for Computational Linguistics. Available at: https://aclanthology.org/P19-1163.pdf (Accessed: 7 February 2024).

Sarno, I. (2023). *What Is a Large Language Model?* [online] knowledge-centre-interpretation.education.ec.europa.eu. Available at: https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model (Accessed: 8 February 2024).

Schakowky, J. (2021), A Bill to provide incentives for and protect whistleblowers under the authority of the Federal Trade Commission, and for other purposes, available at: https://schakowsky.house.gov/sites/evo-subsites/schakowsky-evo.house.gov/files/SCHAKO_082_xml.pdf (Accessed: 7 February 2024).

Schwartz, M. (2008). The Trolls Among Us. *The New York Times*. [online] 3 Aug. Available at: www.nytimes.com/2008/08/03/magazine/03trolls-t.html (Accessed: 7 February 2024).

Shavit, Y. et al. (2023). *Practices for Governing Agentic AI Systems.* Available at: https://openai.com/research/practices-for-governing-agentic-ai-systems (Accessed 5 February 2024).

Shorten, C. and Khoshgoftaar, T.M. (2019). *A survey on Image Data Augmentation for Deep Learning.* Journal of Big Data, [online] 6(1). doi:https://doi.org/10.1186/s40537-019-0197-0. (Accessed: 14 February 2024).

Sneha Solanki (2024). *What is criminal liability? Definition and resources for defense attorneys*. [online] Thomson Reuters Law Blog. Available at: https://legal.thomsonreuters.com/blog/what-is-criminal-liability/ (Accessed: 9 February 2024).

Sijbrandij, S. (2023), *AI weights are not open "source"*, Open Core Ventures, available at: https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/ (Accessed: 9 February 2024).

Silberg, J. and Manyika J. (2019). *Notes from the AI frontier: Tackling bias in AI (and humans)*. McKinsey Global Institute. Available at: www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans (Accessed: 7 February 2024).

Simchon, A., Edwards, M. and Lewandowsky, S. (2024). *The persuasive effects of political microtargeting in the age of generative AI.* Available at: https://academic.oup.com/pnasnexus/article/3/2/pgae035/7591134 (Accessed: 7 February 2024).

South African Competition Commission (2023), *Final Terms of Reference (ToR) for the Media and Digital Platforms Market Inquiry*, Government Gazette No. 49309. Available at: www.gov.za/sites/default/files/gcis_document/202309/49309gon3880.pdf (Accessed: 7 February 2024).

Stanford University Human-Centered Artificial Intelligence (2023). *Artificial Intelligence Index Report 2023 Introduction to the AI Index Report 2023*. Available at: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Accessed: February 7 2024).

Stepanov, A. and Gupta, A. (2021). *Reducing Political Content in News Feed.* Meta. Available at: https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/ (Accessed: 7 February 2024).

Stray, J., Iyer, R., and Puig Larrauri, H. (2023). T*he Algorithmic Management of Polarization and Violence on Social Media.* Knight First Amendment Institute at Columbia University. Available at: https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media (Accessed: 7 February 2024).

Tanenbaum,W., Song, K. and Malek, L. (2022), *Theories of AI liability: It's still about the human element*, Reuters, available at: www.reuters.com/legal/litigation/theories-ai-liability-its-still-about-human-element-2022-09-20/ (Accessed: 7 February 2024).

The Bletchley Declaration by Countries Attending the AI Safety Summit. 1-2 November 2023. Available at: www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 (Accessed: 7 February 2024).

The Global Partnership on Artificial Intelligence (2020). *About - GPAI*. Available at: https://gpai.ai/about/ (Accessed: 7 February 2024).

The National Congress of Brazil (2023). *Bill 2338/2023 'Brazilian Artificial Intelligence Act'*. Available at: https://mcusercontent.com/af97527c75cf28e5d17467eaa/files/248d109f-eeef-7496-4df1-12d29affb522/PL_23382023_Senado_ENG_VF.pdf (Accessed: 7 February 2024).

The United States (1996). *Communications Decency Act, Section 230 of the Telecommunications Act of 1996. 47 U.S.C. § 230*. Available at: www.govinfo.gov/content/pkg/USCODE-2021-title47/pdf/USCODE-2021-title47-chap5-subchapII-partI-sec230.pdf (Accessed: 8 February 2024).

The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Available at: www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (Accessed: 7 February 2024).

Turing (n.d.). *Fine-Tuning LLMs: Overview, Methods & Best Practices*. [online] Available at: www.turing.com/resources/finetuning-large-language-models#what-is-fine-tuning (Accessed: 9 February 2024).

UK Government (2023), *Emerging Processed for AI safety,* available at: https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf (Accessed: 8 February 2024).

UNCTAD (2023), UN Digital Government Awards celebrate excellence in online public services, available at: https://unctad.org/news/un-digital-government-awards-celebrate-excellence-online-public-services (Accessed: 7 February 2024).

UNDP (n.d.) Improving Procurement Transparency, available at: www.undp.org/asia-pacific/fairbiz/improving-procurement-transparency (Accessed: 7 February 2024).

UNESCO (2023). *Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence.* Available at: www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence (Accessed: 7 February 2024).

UNESCO (2023), *Multilevel and Meaningful Transparency in Algorithmic Systems: Developing Concrete Criteria to Guide Institutional and Legal Reforms,* available at: www.unesco.org/en/articles/multilevel-and-meaningful-transparency-algorithmic-systems-developing-concrete-criteria-guide (Accessed: 7 February 2024).

UNESCO (2022). *Windhoek Statement on Artificial Intelligence in Southern Africa.* Available at: https://unesdoc.unesco.org/ark:/48223/pf0000383197 (Accessed: 7 February 2024).

UNESCO (2022), *K-12 AI curricula: A mapping of government-endorsed AI curricula*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000380602 (Accessed on 8 February 2024).

UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence.* Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: 7 February 2024).

UNESCO (2020). Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training. Available at https://en.unesco.org/fightfakenews (Accessed: 8 February 2024).

United Nations (2001). *Universal Declaration on Cultural Diversity.* Available at: www.un.org/en/about-us/universal-declaration-of-human-rights (Accessed: 7 February 2024).

United Nations (1948). *Universal Declaration of Human Rights*. Available at: www.un.org/en/about-us/universal-declaration-of-human-rights (Accessed: 7 February 2024).

United Nations (n.d.), *The Foundation of International Human Rights Law*, available at: www.un.org/en/about-us/udhr/foundation-of-international-human-rights-law (Accessed on 8 February 2024).

United Nations (n.d.). *Civil society*. [online] United Nations. Available at: www.un.org/en/civil-society/page/about-us (Accessed: 8 February 2024).

United Nations Conference on Trade and Development. Data Protection and Privacy Legislation Worldwide. Available at: https://unctad.org/page/data-protection-and-privacy-legislation-worldwide (Accessed: February 7 2024).

United Nations General Assembly (2023). *Promotion and Protection of Human Rights in the Context of Digital Technologies.* Available at: https://digitallibrary.un.org/record/4032837?ln=fr (Accessed: 7 February 2024).

United Nations Secretary-General's AI Advisory Body (2023). *Interim Report: Governing AI for Humanity*. Available at: www.un.org/techenvoy/sites/www.un.org.techenvoy/files/ai_advisory_body_interim_report.pdf (Accessed: February 7 2024).

US Department of Labor (n.d.), Retaliation, available at www.whistleblowers.gov/know_your_rights (Accessed: 9 February 2024).

US Department of Labor (n.d.), *How to Create an Anti-Retaliation Program,* available at www.whistleblowers.gov/antiretaliation (Accessed: 9 February 2024)

US Office of Cyber and Infrastructure Analysis (2018). Social Media Bots Overview. Available at: https://niccs.cisa.gov/sites/default/files/documents/pdf/ncsam_socialmediabotsoverview_508.pdf?trackDocs=ncsam_socialmediabotsoverview_508.pdf (Accessed: 8 February 2024).

Webb, A. (2019), *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*, PublicAffairs New York

Web Foundation et al; (2018), Universal Service and Access Funds: An Untapped Resource to Close the Gender Digital Divide, available at: https://webfoundation.org/docs/2018/03/Using-USAFs-to-Close-the-Gender-Digital-Divide-in-Africa.pdf (Accessed: 9 February 2024).

Wendy Hall, D. and Pesenti, J (2017). *Growing the Artificial Intelligence Industry in the UK*. Available at: https://assets.publishing.service.gov.uk/media/5a824465e5274a2e87dc2079/Growing_the_artificial_intelligence_industry_in_the_UK.pdf (Accessed: 7 February 2024).

Wheeler, T. (2023). *The three challenges of AI regulation.* Brookings. Available at: www.brookings.edu/articles/the-three-challenges-of-ai-regulation/ (Accessed: 7 February 2024).

World Economic Forum (2024). *The Global Risks Report 2024*. Available at: www.weforum.org/publications/global-risks-report-2024/ (Accessed: 7 February 2024).

Yeung, K. (2019, Responsibility and AI, Council of Europe study, available at: https://rm.coe.int/responsability-and-ai-en/168097d9c5 (Accessed: 8 February 2024).

Yu, N. et al (2022). *Artificial Fingerprinting for Generative Models*: *Rooting Deepfake Attribution in Training Data.* Available at: https://arxiv.org/abs/2007.08457 (Accessed: 7 February 2024).

Yun Chee, F (2024). *EU lawmakers ratify political deal on artificial intelligence rules.* Reuters. Available at: www.reuters.com/technology/eu-lawmakers-back-political-deal-artificial-intelligence-rules-2024-02-13/ (Accessed: 13 February 2024)

Zewe, A. (2022), *In machine learning, synthetic data can offer real performance improvements, MIT News.* Available at: https://news.mit.edu/2022/synthetic-data-ai-improvements-1103 (Accessed: 7 February 2024).

Zhang, H. et al. (2023). *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models.* Available at: https://arxiv.org/abs/2311.04378 (Accessed: 15 February 2024).

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M. and Lin, M. (n.d.). *A Recipe for Watermarking Diffusion Models.* Available at: https://arxiv.org/pdf/2303.10137.pdf (Accessed: 7 February 2024).

Zhou, J., Zhang, Y., Luo, Q., Parker, A.G. and Munmun De Choudhury (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.* Available at: https://jiaweizhou.me/assets/chi23_ai_misinfo.pdf (Accessed: 7 February 2024).

Supported by

FORD FOUNDATION

Luminate
Building stronger societies

NED
NATIONAL ENDOWMENT FOR DEMOCRACY
SUPPORTING FREEDOM AROUND THE WORLD

Forum on
Information
& Democracy