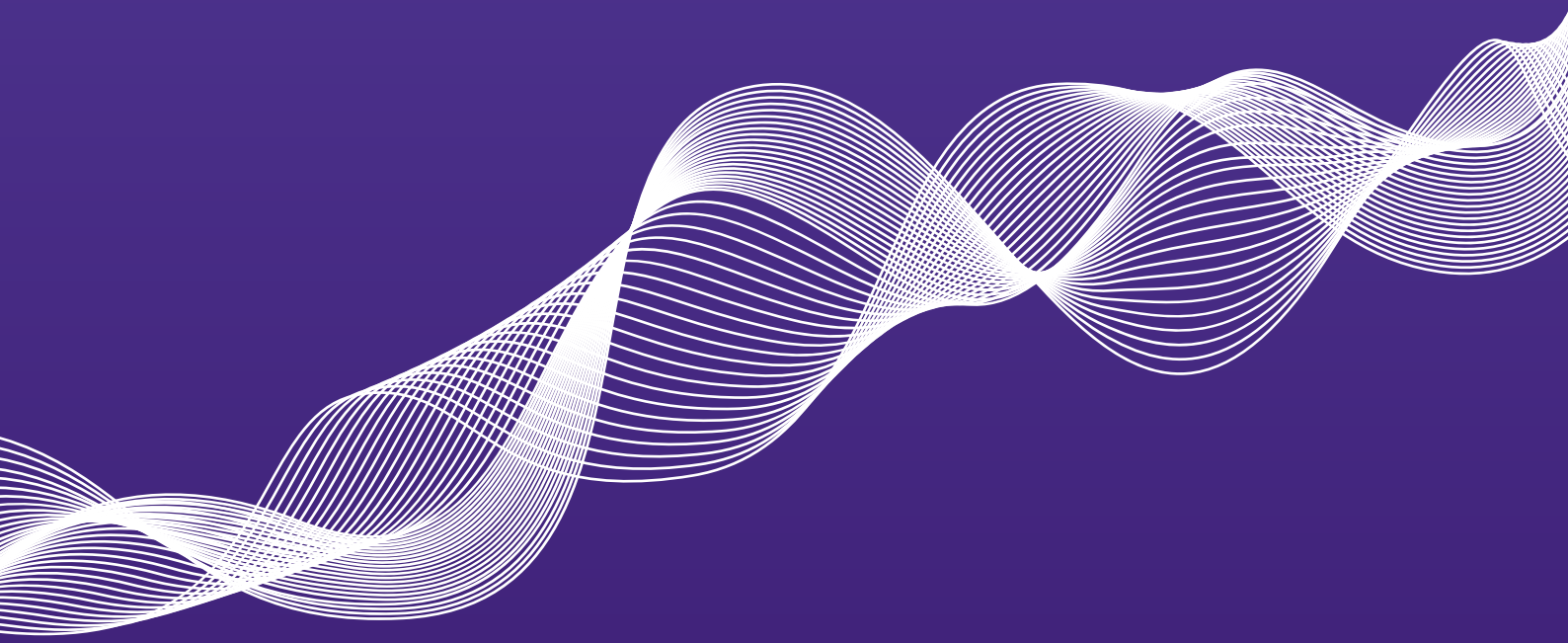
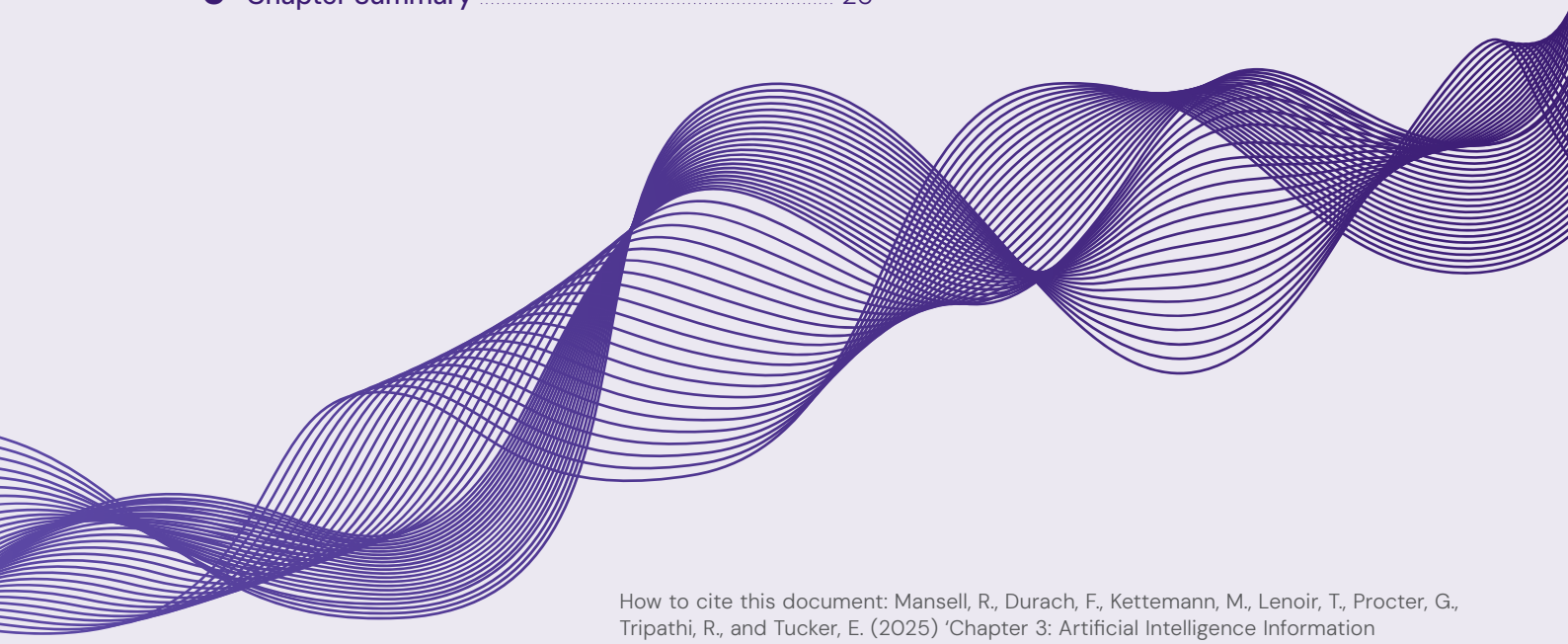
OBSERVATORY ON INFORMATION AND DEMOCRACY

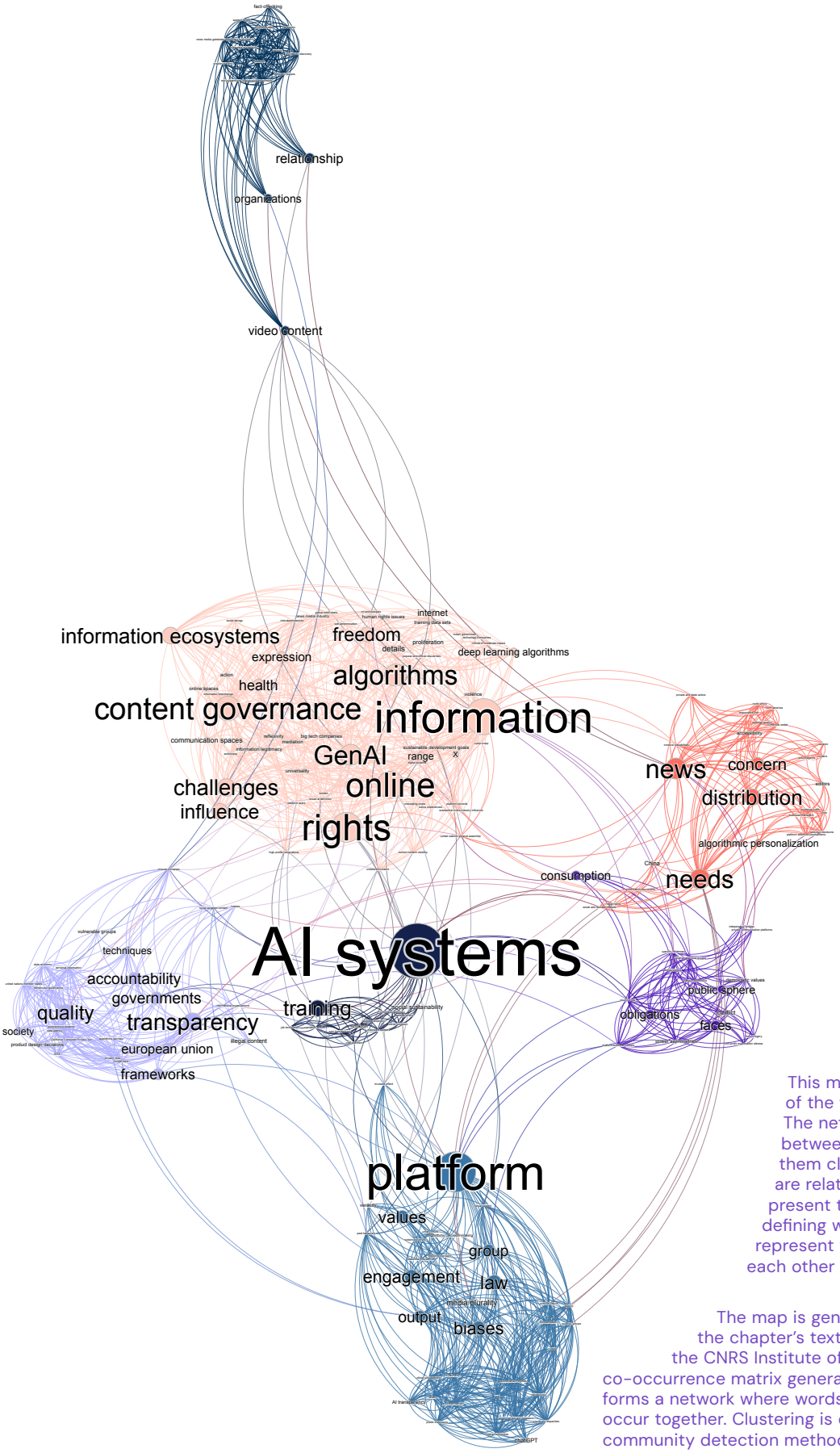
ARTIFICIAL INTELLIGENCE, INFORMATION ECOSYSTEMS AND DEMOCRACY



An initiative by: FORUM ON INFORMATION AND DEMOCRACY



How to cite this document: Mansell, R., Durach, F., Kettemann, M., Lenoir, T., Procter, G., Tripathi, R., and Tucker, E. (2025) 'Chapter 3: Artificial Intelligence Information Ecosystems and Democracy' in Information Ecosystems and Troubled Democracy: A Global Synthesis of the State of Knowledge on New Media, AI and Data Governance. International Observatory on Information and Democracy. Paris.



This map represents a statistical summary of the thematic content of this chapter. The network graph represents relations between the words in the chapter, placing them closer to each other the more they are related. The bigger the node, the more present the word is, signalling its role in defining what the report is about. The colors represent words that are closely related to each other and can be interpreted as a topic.

The map is generated by the OID on the basis of the chapter's text using GarganText – developed by the CNRS Institute of Complex Systems. Starting from a co-occurrence matrix generated from chapter's text, GarganText forms a network where words are connected if they are likely to occur together. Clustering is conducted based on the Louvain community detection method, and the visualization is generated using the Force Atlas 2 algorithm.

Link to the interactive map here

This chapter examines research on the properties of AI systems (specifically machine learning algorithms) and how they are embedded in online content governance systems. It is essential to understand these systems if violations of human rights are to be reduced and flows of mis- and disinformation are not to become an even greater threat to information integrity and to the health of information ecosystems.

The research synthesis focuses on:

- **How is 'artificial intelligence' (AI) defined, and what are the relationships between AI systems development and internationally protected human rights?** The chapter explores whether new rights are needed as AI systems become widely used, and examines the challenges presented by biases in the inputs and outputs of large language models (LLMs). The implications of AI systems for fundamental rights, including freedom of expression and information, privacy and democratic participation, are addressed.
- **What impact do AI systems and content governance, including content generation and content moderation and curation, have on information integrity?** Attention is given to the technologies used for content governance. The use of generative AI (GenAI) by mis- and disinformation actors is also discussed, together with assessments of approaches to countering this type of information and the impacts of generative AI and algorithmic content curation systems on the news media industry.
- **What are the interdependencies between AI systems development, the use of automated tools and democratic processes?** The consequences are discussed, including the influence on debate in the public sphere, the impacts on societal resilience and social sustainability and on environmental sustainability.

The chapter provides a comprehensive assessment of research in these areas, highlighting both the benefits and risks to the health of information ecosystems.

Further discussion of AI systems occurs in later chapters. In Chapters 6 and 7, approaches to AI systems governance that are being put into place by governments, tech companies and not-for-profit organizations are examined. Chapter 8 turns to why the increasing dependency on AI systems and data extraction and processing produces discriminatory outcomes and to strategies aimed at reimagining and practicing alternative approaches to data governance.[1]

[1] For background on AI systems governance, see Bullock *et al.* (2022); Gunkel (2024); Paul *et al.* (2024); Quintavalla & Temperman (2023). For a review of advances in research on generative artificial intelligence (GenAI), including challenges and threats, see Bontcheva *et al.* (2024). See Appendix: Methodology for details of literature review process.

1 Introduction

Humans are social beings. They communicate to achieve common goals, based on convictions they develop through information they receive and share. Democratic decision-making processes cannot function without public discussion of questions of general interest, sharing of ideas and debate about proposed courses of action and past decisions. These processes have become heavily digitalized (i.e., taking place in online spaces) and mediatized (taking place in, and under, the rules, practices and algorithmic systems of privately owned communication spaces). These spaces – information ecosystems – have rules, just as offline spaces do. In offline public spaces, laws set by states and enforced by executive power define the rules for public debate. In online settings, the rules under which communication takes place are set primarily by private actors, such as the owners of the digital platforms in which they take place, within the limits of what the laws allow. These actors enforce their communication rules through systems for content moderation that determine if the content is in keeping with the rules, and curation that decides how to direct the content to platform users. The more platforms seek to automate these systems through the use of artificial intelligence (AI),[2] the more they impact online communication processes and, ultimately, influence democratic discourses and democracy. The integrity of information ecosystems therefore depends on an environment that favors transparency and accountability.[3]

No single definition of 'AI' is accepted by all.[4] The European Union's Artificial Intelligence Act of 2024, for example, defines AI systems as:

> A machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[5]

The OECD definition is similar.[6] Neither of these definitions claims that AI systems emulate human intelligence. Instead, the focus is on functional capabilities that derive from using machine learning (ML) algorithms that work by identifying patterns in data. This interpretation is reinforced by a report prepared for a European Commission Joint Research Centre:

> AI is a generic term that refers to any machine or algorithm that is capable of observing its environment, learning, and based on the knowledge and experience gained, taking intelligent action or proposing decisions. There are many different technologies that fall under this broad AI definition. At the moment, ML [machine learning] techniques are the most widely used.[7]

This definition is interesting because it makes explicit the technologies – that is, algorithms, ML – that constitute AI systems and that other definitions gloss over. It is a reminder of the need to 'look under the hood', to challenge 'the thingness of AI and its status as a stable and agential entity… To let the term pass is to miss the opportunity to trace its sources of power and to demystify its referents'.[8] It is therefore important to engage in a critical discussion on what 'AI' is. However, two factors

[2] *The Eye of the Master* presents a social history of AI systems, emphasizing that they are not 'intelligent' and that work in this field has been motivated historically by interests in labor saving and surveillance (Pasquinelli, 2023). There are many warnings about the inherent problems in anthropomorphizing AI systems (Floridi & Nobre, 2024). There are suggestions for a new glossary of terms, for example, 'systems for statistical propositions', to describe large language models (LLMs) to support discussion of the benefits and harms of technological advances more transparently (Frau-Meigs, 2024b). In the field of political communication, for example, 'AI' has been defined as 'the tangible real-world capability of non-human machines or artificial entities to perform, task solve, communicate, interact, and act logically as it occurs with biological humans' (Gil de Zúñiga *et al.*, 2023, p. 2), supported by the Spanish National Research Council (CSIC, Consejo Superior de Investigaciones Científicas). Activists and critical scholars emphasize the importance of focusing not just on technology, but also on politics, power structures, cultural narratives and public perceptions (PublicSpaces International, 2024; Verdegem, 2021).

[3] Nowotny (2021); Puddephatt (2021).

[4] Samoili *et al.* (2020).

[5] EC (2024c, Article 3(1)).

[6] An AI system is 'a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment' (OECD, 2022c, p. 7).

[7] Annoni *et al.* (2018, p. 18).

[8] Suchman (2023, p. 1).

make the use of the term 'AI' in this report difficult to avoid; first, the proprietary nature of many systems means that details of the technologies used are often not disclosed; and second, 'AI' is widely used, not only in the research literature, but also in both popular and official discourses.

So-called 'generative AI' (GenAI) refers to a broad category of ML systems that are capable of synthesizing content. They are typically trained on very large data sets and can generate content – synthetic media – in the form of text, images and video that may often be difficult to distinguish in terms of quality from human-generated content. Among the various examples of GenAI systems, large language models (LLMs) are the best known. Despite being classified as GenAI, however, LLMs are simply statistical models of language use. While systems that use LLMs, such as chatbots, can produce very plausible responses to queries, this should not be mistaken for *natural language understanding*. An LLM, then, is: 'a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: *a stochastic parrot*'.[9]

LLMs first achieved public attention in November 2022 with the announcement of ChatGPT by OpenAI, and are already being used in ways that have significant implications for the public's experiences of information ecosystems and the content that diffuses through them. These include, for example, to create and moderate content such as hate speech; to create realistic 'deepfakes', but also to detect them; and to promote, but also to fight, mis- and disinformation.[10] And as the realism of 'deepfakes' increases, their detection becomes correspondingly harder.[11]

In its 2024 report, *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, the Forum on Information and Democracy said that AI systems, particularly GenAI systems, are 'revolutionizing the way we create information across various mediums, including text, audio, images and video, presenting both challenges and opportunities'.[12] Gaining democratic control of AI systems requires effective accountability structures for the whole AI systems lifecycle, which the OECD defines as setting objectives and the functional specification, building a model to meet the specification, and its verification and validation as well as its deployment, operation and monitoring.[13]

It will be clear that there is not *an* AI; rather there are different ML technologies, instances of which may be involved in processes related to information retrieval, synthesis, presentation and governance. ML technologies vary widely, ranging from relatively simple algorithms executing tasks (such as filtering for specific words) to deep learning algorithms (that can be trained to assess the likelihood of content having been authored by an inauthentic actor, such as a disinformation bot).[14]

Embedding AI systems within information ecosystems impacts on content production (synthesizing text, images and video), moderation (deciding if content violates regulations) and consumption (deciding on the content's audience). It is therefore unsurprising that concerns have been raised about the potential for these systems to shape public discourse and, moreover, to do so in ways that may have significant implications for societal cohesion and resilience.[15] Questions about the use of 'AI' in information ecosystems cannot be settled on technical criteria alone, but must address a much broader range of issues, including legal (e.g., does their use discriminate against certain groups?) and societal (e.g., does their use reduce the diversity of information available to publics?).[16]

[9] *Bender et al.* (2021, p. 617), supported by the National Science Foundation (NSF), US.
[10] See Bonfanti (2020), Real Instituto Elcano, independent think tank; Kertysova (2018); Spitale *et al.* (2023).
[11] Ghosal *et al.* (2023). The prevalence of 'deepfakes' and other types of mis- and disinformation is discussed in Chapter 5, and efforts to combat them are discussed in Section 3.5 of this Chapter, and more extensively in Chapter 7.
[12] Forum on Information and Democracy (2024a, p. 18).
[13] OECD (2023).
[14] Veale *et al.* (2023).
[15] De Gregorio & Stremlau (2023), supported in part by the European Commission.

As the capabilities of AI systems continue to advance and find application within information ecosystems, it is anticipated 'that algorithmic moderation and regulation will become more and more seamlessly integrated into our social lives'.[17] As this process progresses, the increased 'consumption and commodification of artificial intelligence applications in daily life', coupled with the 'extensive trust and reliance on these technologies in public and private sectors', makes it essential to confront important rule of law questions.[18] The many different ways in which these and other questions may be answered should act as a timely reminder that how new technologies become embedded within people's everyday lives is neither inevitable nor identical in different countries and regions, and can be shaped and influenced through normative choices based on ethical values and societal (and international) goals to be pursued (or not), as well as the experience and outlook of people in different regions and countries.

Discussions about how to ensure the health of information ecosystems that increasingly depend on AI for their day-to-day function need to be as inclusive as possible. While the Global North deals with the effects of the fast-growing pace of technological change on information ecosystems, the Global Majority World struggles with issues such as access to the internet, inequalities in investment in online safety and content governance resources, poor infrastructure and weak technology literacy levels.[19] This means that some parts of the world are excluded from experiencing the benefits of AI systems (as well as other components of the digital infrastructure). As Kenichiro Natsume, Assistant Director-General at the World Intellectual Property Organization (WIPO), pointed out, '[the] 2.6 billion [unconnected] people [who] are not part of the digital world … are [also] not part of the AI world'.[20] Exclusion from the internet keeps this population, which is disproportionately located in the Global Majority World, from accessing online information, and also from using AI tools, including GenAI.[21] This does not mean these populations are unaffected by 'AI divides' since they are recipients of information that circulates by other means. Even when internet connectivity is achieved and affordable, the terms and conditions of online information access are skewed and shaped by big tech companies and by communication infrastructure providers that influence what information can be accessed, and which information is amplified by AI systems use and algorithms for those who are connected.[22]

2 AI Systems and Human Rights

This section examines how human rights apply in the digital age, the problems created for fairness as a result of algorithmic biases, the importance of freedom of expression and information as well as privacy protection in considering the impacts of AI systems developments, and the impact on participatory rights as a result of the use of AI systems to manipulate information.

2.1 NEW TECHNOLOGIES – BUT NO NEW RIGHTS

Human dignity serves as the cornerstone of human rights. Thirty years ago, the guiding principles of the Vienna Declaration on human rights emphasized the indivisibility, universality, interrelatedness, and mutually dependent and reinforcing nature of all human rights.[23] Predating this, the *Universal Declaration of Human Rights* (UDHR) of 1948

[16] Katzenbach (2021). This is especially so when companies such as OpenAI put 'shiny products' above safety, as claimed by researchers who have since left the company (Milmo, 2024).
[17] Katzenbach (2021, p. 6).
[18] De Gregorio (2023, p. 1).
[19] De Gregorio & Stremlau (2023), supported in part by the European Commission.
[20] Quoted in Vanoli (2024).
[21] Fendji (2024). Some have limited access by sharing internet access accounts, but others have no internet access, due to absent or costly infrastructure. See Heeks (2022), supported by the Economic and Social Research Council (ESRC), UK; Mutsvairo & Ragnedda (2019); Okolo (2023).
[22] This issue is discussed further in Section 4.1, Chapter 6.
[23] OHCHR (1993).

committed states to the 'promotion of universal respect for and observance of human rights and fundamental freedoms', declaring these rights a 'common standard of achievement for all peoples and nations'.[24]

Human rights are fully applicable in the age of digital transformations, although much work is needed to uphold them in practice. As the then-United Nations High Commissioner for Human Rights, Michelle Bachelet, concluded in a speech in 2019, technology change does not necessitate new human rights conventions, but rather: 'adapting the way we use institutions and processes… We can protect rights effectively only if we constantly fine-tune our processes to find the right mix of interventions'.[25]

All societal actors have human rights obligations, albeit to differing degrees. Private online communication platforms have duties under the so-called Ruggie Principles, the *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*.[26] Private entities need to protect, respect and provide remedies for violations of human rights, under the overall control of states. Following international human rights law, states have to respect, protect and ensure these rights for anyone within their control or jurisdiction,[27] although in the absence of regulation, these duties are not necessarily binding on all actors.

While digital platforms tend to frame their mission in human rights terms, such as 'giving people a voice' or 'protecting expression', the focus of research has been primarily on potential human rights violations by governments and less on areas where platform business models might negatively impact user rights.[28] In the light of a reluctance to commit to substantial transparency obligations over the last decades, regional and national approaches have emerged to apply human rights obligations more directly to platforms.[29] The first 'big picture' approach can be seen in the European Union's Digital Strategy, which attempts to curtail the influence of large digital companies by imposing obligations on them that mitigate the negative effects of online communication and, at the same time, promote the implementation of fundamental rights.[30] The European Union has emerged as a key international norm-maker for the digital arena, sometimes referred to as the 'Brussels Effect'.[31] Legislation, including the AI Act of 2024, provides some substantive obligations, but through stringent transparency and compliance obligations.[32] Selected human rights issues that arise in the context of automated content governance and that impact on democratic decision-making processes are outlined below.

United Nations initiatives, such as a March 2024 General Assembly Resolution, show how there is awareness of technology's role in both contributing to disruptive change and having the potential to build bridges within and between countries. The Resolution emphasizes that trustworthy AI systems for sustainable development should be promoted globally in line with existing human rights obligations.[33] By September 2024, AI systems had been positioned with other frontier technologies as a means to 'turbocharge development', securing a place as Objective 5 of the United Nations' Global Digital Compact, which emphasizes the need for a 'balanced, inclusive and risk-based approach to the governance of artificial intelligence (AI)'.[34]

[24] UN (1948, preamble).
[25] Bachelet (2019).
[26] Ruggie (2011).
[27] Fischer-Lescano (2016), funded by the European Research Council (ERC).
[28] Jørgensen (2017); Kettemann & Schulz (2023).
[29] Müller & Kettemann (2024).
[30] EC (2022b).
[31] Bradford (2020).
[32] EC (2024c); Müller & Kettemann (2024); Werthner *et al.* (2024). Governance arrangements for these technologies are discussed in Chapters 6 and 7.
[33] UN (2024c), adopted by the UN General Assembly on 21 March 2024.
[34] UN (2024b, pp. 41, 52).

2.2 ALGORITHMIC BIAS AND FAIRNESS

Algorithmic bias involves systematic errors within AI systems that lead to unfair results.[35] Unfairness can be understood as privileging, without adequate reasons, members of one group over another. When used in settings where automated decisions impact individual or collective rights or values, these biases can lead to unfair and untransparent outcomes, not least because of economic incentives to favor results consistent with corporate interests.[36] Algorithmic decision-making in areas such as employment, law enforcement and lending can disproportionately negatively affect marginalized communities, and contribute to their exclusion from participation in democratic processes or the full enjoyment of their rights.[37] Research predominantly in the Global North, but also in the Global Majority World, reveals how algorithmic bias can lead to decisions by law enforcement authorities that disproportionately penalize minority ethnic groups and immigrant communities.[38]

Biases arise from various factors linked to how, by whom and in which institutional or organizational setting an AI system is developed, particularly regarding the data used to train it – for example, when training data is incomplete or contains historical prejudices or assumptions that are then replicated: if, in the text on which an LLM is trained doctors are primarily described as male, then answers generated by the LLM will replicate this.[39] Similar replication of stereotypes has been shown to happen in image-generating LLMs. Even AI systems trained using what is believed to be unbiased data may produce biased outputs, since a lack of transparency in how their outputs are produced may make it difficult to exercise effective oversight over their performance.[40] The personalization algorithms used on social media platforms to decide what content users are exposed to exploit the data users create when they interact with content. Once 'datafied'[41] in this way, AI algorithms can be used to model user behavior, and the model can then be applied in ways that are biased towards the interests of platforms, leading to the promotion of content that maximizes user engagement at the expense of quality and veracity.[42] These are all consequences of the way that LLMs synthesize their training data to produce outputs based on statistical prevalence – reducing the diversity of inputs into the specificity of a single output. In addition, LLMs may be trained on synthetic data, that is, 'data that mimic and substitute empirical observations without directly corresponding to real-world phenomena'.[43] Critical assessments of the use of such data may be helpful in protecting privacy and improving data sets that have a representational link to the 'real-world', for example, addressing biases, but when developed by artificial neural networks this does not provide a means of explaining why a given output has been generated.

Algorithmic fairness refers to the aspiration of creating and implementing AI systems that do not discriminate or bias against specific persons or groups based on protected characteristics, such as race, gender or ethnicity.[44] Fair AI algorithms would make decisions without favoring one individual or group over another.[45] To achieve this, attempts are now being made to increase the quality of training data sets. IBM launched a Diversity in Faces data set to help overcome specific biases in facial recognition technology.[46] This data set includes a million images of faces annotated with details that provide a broad representation of human faces, such as age, gender, skin tone and facial

[35] Hasimi & Poniszewska-Marańda (2024); see further discussion of fairness Sections 2 & 3, Chapter 4 and in Chapter 8.
[36] The biases of personalization systems and search engines have been recognized in the literature and demonstrated empirically for at least a decade (Eubanks, 2018; Rieder & Sire, 2014).
[37] Baecker *et al.* (2023).
[38] Chouliaraki & Georgiou (2022); Gurumurthy & Chami (2019).
[39] Belenguer (2022).
[40] Pollicino & De Gregorio (2022).
[41] 'Datafied' means turning a previously computationally invisible activity into data, and is a term used especially in the literature that is critical of the datafication of the lives of human beings (van Dijck, 2014).
[42] *Pfeiffer et al.* (2023), funded by Projekt DEAL, Alliance of Science Organizations, Germany.
[43] Offenhuber (2024, p. 1), and for a discussion of a variety of types of synthetic data and their implications.
[44] Ferrara (2024a); Johnson (2023).
[45] Hall & Ellis (2023).
[46] Smith (2019); the author was an IBM employee.

features drawn from many different countries and cultures. By using this data set, developers can train facial recognition systems that are less likely to reproduce stereotypes regarding certain groups.[47] This approach is based, however, on the premise that greater diversity will reduce the prevalence of bias, but can be limited by the unavailability of more diverse training data. The challenges around guaranteeing fairness will increase as AI progressively becomes enmeshed in the processes that define the social conditions in which meaning is produced. These, in turn, are dependent on the level of trust in them, their prevalence and institutional roles.[48]

Diversity in training data is expected to contribute to mitigating the risks of bias in AI systems that use these models. Diversity in development teams can offer a variety of perspectives that challenge conventional norms and biases that may be overlooked in more homogenous teams. The setup of development teams – and those working on AI ethics generally – is substantially linked to product design decisions.[49] Microsoft has embraced this strategy through its Inclusive Design Initiative, which employs people with diverse backgrounds (including disabilities) to design and test new products.[50] Evidence of the effectiveness of such corporate diversity strategies is inconclusive, and in some cases no direct association is found between the socio-demographic diversity of AI systems developers and AI systems output biases. The viewpoint diversity of those holding ML, coding or data analyst jobs is found to play a much stronger role based on a relatively small-scale study.[51] Various forms of discrimination are likely to persist in the prevailing culture, which is likely to be encouraged if its leadership is skewed to favor certain groups, as illustrated by high-profile resignations from some of the big tech companies.

In many countries of the Global Majority World, AI systems development and deployment are at a 'nascent stage', potentially allowing countries to design robust anti-discrimination rules before broad uptake.[52] For the Global Majority World, questions about 'human rights, democracy and autonomy in the countries of the majority world are not trivial'.[53] For example, the development of fair AI systems may be hindered by the limited availability of training data in many Global Majority World languages.[54]

2.3 FREEDOM OF EXPRESSION AND INFORMATION

Freedom of expression is a 'cornerstone' for the formation of democratic societies, and as such is protected by all human rights instruments, including Article 19 of the UDHR and Article 19 of the *International Covenant on Civil and Political Rights* (ICCPR), and all regional human rights conventions.[55] This right includes the freedom to express and hold one's own opinions, to impart information, to seek and receive information and, implicitly, freedom of media expression. Given the technological realities of online communication, the right to freedom of expression is implicated in other rights such as the right to health (seeking and imparting health-related information) and to education (seeking and imparting information related to education, attending classes and research papers).

AI systems allow for much easier access to online communication spaces and information interchange, but also impact what information can be seen.[56] All platforms use AI systems to govern online communication and optimize user engagement.[57] There is thus a substantial impact, across information ecosystems, of these content governance systems on freedom of expression.[58]

[47] Wiggers (2019).
[48] Pfeiffer *et al.* (2023).
[49] Martin (2022).
[50] Microsoft (2023).
[51] Chi *et al.* (2021); Harris (2023); Park (2024).
[52] Gurumurthy & Chami (2019, p. 9).
[53] Ricaurte (2022, p. 732).
[54] Ricaurte (2022), citing Horowitz (2021); more recently, see HRW (2023).
[55] UN (1948, 1966).
[56] Dias Oliva (2020).
[57] Gillespie (2020); Longo *et al.* (2024).
[58] De Gregorio & Dunn (2023).

Figure 3.1
Illustration of user engagement

Source: Pixabay

Information enables individuals to make educated judgments by helping them become acquainted with facts (see Figure 3.1) and societal issues.[59] It is a crucial component of individual liberty. Nevertheless, people's ability to obtain and interpret information can be restricted when they encounter mis- or disinformation or biased content.[60] Another challenge people face is a lack of reliable, accurate information, a problem sometimes compounded by information overload, and worsened when there is a decline in global trust in news, which is associated with the prevalence of online mis- and disinformation, as discussed in Chapter 2.[61]

Focusing mainly on tweaking content governance practices and systems ignores the underlying causes of social discord and distrust that give rise to polarized public opinion. Some argue that a focus on the 'public worthiness' of information, rather than on information 'disorder', can reveal the complex elements of visibility, access, reflexivity, mediation, influence and information legitimacy. Better insight into how these can combine in different ways to foster new imaginings of publicness could enable democracy to flourish.[62]

2.4 PRIVACY PROTECTION

AI systems present significant challenges to people's right to data protection and privacy. The comprehensive collection and analysis of data by these systems leads to the development of multidata, points-based profiles of individuals, often without their explicit consent, which can separately and in aggregate violate their right to privacy.[63] For example, Meta has said it is extending the jurisdictions in which it collects public data to train its models beyond the United States, although Data Protection Authorities in the European Union and Brazil and beyond have sought to stop this practice. In other countries, such as the United Kingdom, after some changes this practice has been deemed a 'legitimate interest' in processing data.[64] Moreover, the way consent is obtained for data collection does not often meet the threshold of being 'informed'. Many users experience consent fatigue, agreeing to privacy policies without understanding the implications.[65]

As most online communication takes place in private communication spaces that are financed through data collection, there is an incentive for platforms that use automated content governance tools to configure them in a way that maximizes data collection. This can lead to interferences with, and violations of, rights to privacy and data protection. These can be addressed to some extent by enforcing existing privacy and data protection laws and international human rights standards that emphasize consent, data minimization and purpose limitation in data processing – for example, Article 12 of the UDHR and Article 17 of the ICCPR protect privacy and personal data.[66]

[59] Masur (2020).
[60] Measurement issues around the scale of mis- and disinformation are discussed in Section 2, Chapter 5, along with issues of public awareness of its prevalence in Section 3, Chapter 5.
[61] Samoilenko & Suvorova (2023). The big tech platform's practices of reducing or amplifying content, for example, of reducing news media or user-generated content, are discussed in Chapter 2 (Table 2.1), Chapters 6 and 7 as a self-regulatory strategy.
[62] Splichal (2022a); Geiß *et al.* (2021, p. 683), supported in part by the Media Authority of North Rhine-Westphalia, Germany.
[63] Bontridder & Poullet (2021).
[64] Forum on Information and Democracy (2024b).
[65] Abdulrauf & Dube (2024); Barocas & Nissenbaum (2014); Richards & Hartzog (2019); Turow *et al.* (2023); Avle (2022), supported by the National Science Foundation (NSF), US.
[66] UN (1948, 1966).

Training an LLM on personal data effectively encodes aspects of this data into the model's parameters. Even where there is no direct retention of the data,[67] the model learns patterns and information during training that may be reconstructed or inferred by analyzing its output. Studies have shown that it is possible to extract specific data points from LLMs through techniques like model inversion or membership inference attacks, where queries to the model can reveal if certain data was used in training.[68] However, this approach may be limited by the fact that it suggests that it is possible to invert outputs to inputs, ignoring that the model combines inputs according to probabilistic weights that are derived from a combination of inputs, rather than linearly from any single input. In addition, models learn and change when model–user interaction or 'user embedding' occurs in addition to learning in response to user text prompts.[69]

There are several potential technical solutions and strategic reforms that can be implemented to address the privacy risks posed by LLMs and other types of AI. These aim to enhance privacy protection, ensure transparency and uphold ethical standards within AI systems development and deployment.[70] Regarding technical solutions, protection from de-anonymization risks can be achieved by using differential privacy methods that add random noise to the data in a way that prevents the identification of any individual from the data set, and these methods have been adopted by companies such as Apple and Google.[71]

Establishing and adhering to ethical standards when developing AI systems is essential to mitigate risks related to privacy, bias and other potential harms. For instance, the Institute of Electrical and Electronics Engineers (IEEE) has proposed an ethical framework for AI and autonomous systems that includes guidelines for prioritizing human well-being, data agency and accountability in AI systems.[72] Similarly, the Partnership on Artificial Intelligence to Benefit People and Society (Partnership on AI), which includes stakeholders from various organizations, promotes best practice in AI development, focusing on fairness, transparency and accountability in an effort to ensure AI systems are used responsibly, although substantial changes in company policies or product priorities have not materialized.[73]

Legislative tools such as the European Union's General Data Protection Regulation (GDPR), its AI Act or the United States' California Consumer Privacy Act (CCPA) provide foundational frameworks for regulating AI systems and data use practices, aiming for comprehensive protection for individuals' privacy.[74] This approach includes strict requirements for transparency and data quality, setting a benchmark for global AI regulations, aiming to empower consumers with more control over the personal information that businesses collect about them – including transparency about data use and the right to delete collected personal data. Research suggests, however, that both the GDPR and CCPA have significant limitations and, in the case of the GDPR, regarding informing data subjects about how their data is being used.[75]

Surveillance is defined as the 'process of observing individuals or groups for a purpose and make inferences/judgements on their behavior'.[76] Its scope and scale have been transformed by 'datafication', that is, the quantification of people's everyday activities in real time by digital platforms.[77] AI systems algorithms can then be used to analyze this data to identify patterns of behavior. The risks

[67] This will depend on whether training requires access to personal data and where this data is stored. In the case of ChatGPT, any additional data must be uploaded to OpenAI's servers and OpenAI retains this data. Some LLMs allow for data to be retained locally.

[68] Jagannatha *et al.* (2021).

[69] Ning *et al.* (2024).

[70] Lepri *et al.* (2018); Yan *et al.* (2024); Ong *et al.* (2024), supported in part by the Wellcome Trust.

[71] Zhao & Chen (2022).

[72] IEEE (2019); Gunkel (2024); see also UNESCO's recommendation on the ethics of AI (2022c).

[73] Borocas *et al.* (2023); Caton & Haas (2020).

[74] EC (2016b, 2024c); Mahler (2022); US State of California (2018). See also Section 4.2, Chapter 6.

[75] Lee (2024); Wulf & Seizov (2022), supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) and Dutch Research Council (NWO, Nederlandse Organisatie voor Wetenschappelijk Onderzoek).

[76] Fontes *et al.* (2022, p. 2).

[77] 'Datafication' refers to the 'transformation of social action into quantified data' for real-time tracking and prediction (van Dijck, 2014, p. 198).

of – and potential remedies to – surveillance in the data economy, sometimes known as 'dataveillance', are discussed further in later chapters.[78]

2.5 DEMOCRACY AND PARTICIPATORY RIGHTS

While United Nations member states are committed to democratic governance, the majority of the world's population live in states that suffer from democratic deficits. A 2023 analysis of the state of democracy globally concluded that it is 'complex, fluid and unequal'.[79] The *Vienna Declaration and Programme of Action* of 1993 clearly links democracy with human rights, urging member states to continuously foster democratic principles to enhance human rights protection.[80] Democracy fundamentally relies on the principles of free, equal, secret and independent elections and democratic decision-making processes. Particularly problematic are targeted mis- and disinformation campaigns that aim to manipulate elections and stir conflict.[81] The manipulation of democratic decision-making processes is facilitated by AI systems. Even non-subliminal techniques can be manipulative, especially for vulnerable groups such as children, if they exploit mental health vulnerabilities, immaturity or lack of digital literacy. As AI systems evolve, the potential for misuse increases. A notable example is OpenAI's Sora, one of several GenAI systems capable of producing video footage from minimal text input.[82]

The utilization of AI systems in political campaigns and election processes leads to concerns about the transparency, accountability and manipulation of democratic decision-making. The electoral landscape faces significant risks from the very tools that enable campaigns to target voters with exceptional accuracy: the capacity to disseminate false information, to manipulate perceptions through microtargeting and to magnify controversial content.[83] The capacity of AI chatbots, for example Microsoft's Bing Chat, was tested over several months during elections in Germany and Switzerland. This GenAI chatbot produced factual errors to queries on election topics with a near 30% error rate.[84] The use of AI systems to personalize content on social media platforms has the potential to sway voters and create divisions in public opinion, affecting individuals' abilities to freely engage in their government and public affairs. Advanced data analytic capabilities have made voter microtargeting significantly more accessible. While this has the capacity to enhance engagement and voting percentages, it also exposes voters to manipulation via hyper-targeted content that can seek to sway their opinions or even discourage them from voting.[85]

3 AI Systems and Content Governance

AI systems deployed by digital platforms manage the visibility and spread of information, mis- and disinformation.[86] This section addresses content generation and governance, that is, content moderation, distribution and amplification; it assesses the impact of AI systems on information ecosystems; and discusses how AI systems are being used by mis- and disinformation actors.

Social media platforms have become vital arenas for public debate, where users gather information, share ideas and form opinions. Content governance systems impact on these processes because they frame the conditions under which content is seen and with whom it is shared.[87] These systems utilize

[78] 'Dataveillance' refers to continuous surveillance using (meta)data (van Dijck, 2014). Surveillance is examined further in Chapters 4 and 8.
[79] International IDEA (2023).
[80] UN OHCHR (1993).
[81] See Section 4.3.3, Chapter 2 for a discussion of the weaponization of information and election manipulation.
[82] Liu *et al.* (2024); one of the authors works with Microsoft Research.
[83] Schippers (2020).
[84] See Helming (2023). The impact of mis- and disinformation on political processes is discussed in Section 3, Chapter 2 as well as in Chapter 5, where AI literacy and capacities to discern accurate from inaccurate information, including the 'hallucinations' generated by AI systems, are discussed.
[85] Michael (2023); the Cambridge Analytica story is discussed in Section 4.3.3, Chapter 2.
[86] Sančanin & Penjišević (2022).

user behavior, previous choices (interest histories) and past interactions to customize content streams, control content visibility and enhance engagement metrics. AI-based content governance systems are intended to reduce the prevalence of undesired content such as mis- or disinformation, including hate speech and propaganda.[88] Importantly, their design, implementation and accountability lie in the hands of the platforms where they are used; these systems, and the governance policies and practices they are intended to support, vary from platform to platform and from jurisdiction to jurisdiction, changing over time, especially with changes in ownership, as illustrated in the case of X/Twitter.[89]

Twitter's transformation under Musk. On his takeover of Twitter in late 2022, Elon Musk announced: 'The reason I acquired Twitter is because it is important to the future of civilization to have a common digital town square, where a wide range of beliefs can be debated in a healthy manner, without resorting to violence'.[90] Not long after, he introduced significant changes to Twitter's content policies and practices, signaled by reinstating some high-profile users who had been banned for violating the platform's misinformation and hateful conduct policies.[91] Changes in practices were inevitable, with the sacking of a large proportion of staff responsible for human rights, AI ethics, trust and safety.[92] X introduced Community Notes, which aim 'to create a better informed world by empowering people on X to collaboratively add context to potentially misleading posts',[93] but retains control over which contributions are approved and made visible to users.

The Australian eSafety Commissioner has criticized X for letting the worst offenders back online, 'while at the same time significantly reducing trust and safety personnel whose job it is to protect users from harm'.[94]

The AI algorithms that drive social media platforms are designed to enhance user engagement by personalizing online experiences, and are, in principle, neutral on the veracity of content. However, if mis- or disinformation content provides the most engagement, the system – if not properly reviewed – will increase the dissemination of such content.[95]

3.1 AI SYSTEMS IN CONTENT GENERATION

The availability and ease of use of GenAI has arguably 'democratized' content production. Making a video used to be the reserve of a privileged few. Without specific detailed technical know-how, users can now create digital content in audio, video or text, using a wide range of apps, and distribute them through digital platforms.[96] With this comes potential 'side effects', which stem from the increase in volume, velocity and potential persuasiveness of problematic content and its decreasing cost of production.[97]

Digital platforms have started to address the challenges of text and speech produced by GenAI, but, in jurisdictions without rules on risk assessment obligations, the internal rules are often vague or inconsistently enforced: 'The driving force is either the misleading and harmful potential or a more compliance-oriented approach in terms of copyright and quality standards of the content'.[98]

[87] Jungherr & Schroeder (2023), funded by the Volkswagen Foundation (Volkswagen Stiftung).
[88] Christodoulou & Iordanou (2021), funded by the European Commission.
[89] Burkart & Huber (2021); see EC (2024d), for demanding that X explain its content moderation compliance with European Union regulations.
[90] York (2022).
[91] Ivanova (2022).
[92] Brewster (2024); eSafety Commissioner (2024a).
[93] X (2024).
[94] eSafety Commissioner (2024b).
[95] Bontridder & Poullet (2021); Ohme *et al.* (2024); Reisach (2021); see also Chapter 2 where audience/user engagement with content and mis- and disinformation research is discussed.
[96] Allen & Weyl (2024); Cooke (2023).
[97] Feuerriegel *et al.* (2023).
[98] Miguel & Krack (2023, p. 3).

Given that audiences find it difficult to distinguish between GenAI and human-produced content,[99] it is important to raise levels of AI literacy and to impose disclosure obligations for AI-generated content or AI-operated accounts.[100] So far, however, despite high-profile cases, there is no evidence that GenAI is *systematically* used as a tool to synthesize politically motivated mis- and disinformation.[101] Even if this is the case, there is no doubt that it is being used with growing pressures on tech companies, prompting them to sign a voluntary accord in early 2024 to prevent AI systems from disrupting elections.[102]

Copyright is a challenging issue for AI-generated content because it applies both to the data used for training and to the generated output. Training AI systems, especially LLMs, often involves ingesting vast amounts of text harvested from the internet, much of which is copyrighted (although some exceptions exist), raising questions about whether this usage constitutes 'fair use' or 'exceptions' to copyright depending on the jurisdiction, or requires explicit permission from rights holders. The output generated by AI systems also sometimes (and inexplicably) reproduces copyrighted material verbatim.[103]

3.2 AI SYSTEMS IN CONTENT MODERATION AND CURATION

AI systems are increasingly used by platforms for implementing content governance guidelines on how content is sourced (or created) and then distributed. Content moderation involves identifying and removing or flagging inappropriate, harmful or illegal content based on predefined criteria. This definition of criteria, the setting up of internal standards and community guidelines is a powerful act, which, coupled with algorithmic content moderation and curation (i.e., governance), gives digital platforms a role that researchers call the 'arbiters of truth'.[104] It is not so much 'truth' that is decided on, however, but what content stays on a platform and what content is given more visibility. Content curation systems then select and organize content that has passed the moderation stage for distribution. These systems are used by platforms to determine who sees what content, often personalizing it by matching against users' preferences, as revealed by their past behaviors.[105] The use of these systems takes place within the framework of existing and new laws shaping platform behavior, including rules for transparency and user rights. No moderation or content curation system is neutral or non-discriminatory. If it did not treat content differently, it would not be doing its job. Certain categories and procedures must be used to structure the content presented to social media users. Choices must be made even if the choice is to present content in chronological order. As a report for UNESCO's regional office in Montevideo put it:

> AI technologies are not neutral; they inherently reflect the values of their developers and the broader development and deployment ecosystem. While they have the potential to enhance accountability in public institutions and their representatives, foster greater participation and pluralism to enrich citizen engagement, and make democracy more inclusive and responsive, they can also amplify autocratic tendencies and be used for potentially malicious and manipulative purposes.[106]

[99] Kreps *et al.* (2022).

[100] AI literacy is discussed in Chapter 5.

[101] Kreps *et al.* (2022); Simon *et al.* (2023).

[102] O'Brien & Swenson (2024). This accord is discussed in the context of the governance of political processes in Chapter 7.

[103] Geiger (2024) discusses a human rights-friendly copyright framework for GenAI, emphasizing the rights of human creators. UNESCO began considering the impact of AI systems on cultural production earlier than the current debate about LLMs (Kulesz, 2018). WIPO states that there is significant legal uncertainty, and answers are likely to vary by jurisdiction (2024). In the European Union, if a work is created by AI, it is not subject to copyright, but there is scope for application of the law if a creator is deemed to have given explicit instructions to an AI application. The AI Act says that text and data-mining operations must receive consent unless they are subject to exemptions – which so far seem to apply – but companies must document their use of data and court proceedings are underway. As of August 2024, the United States does not offer copyright protection to creations produced by GenAI, and it is not clear what liability OpenAI and other firms have for scaping data to train LLMs. Legislation is being presented to Congress, but none has succeeded in becoming law. The issues in this area relating to 'fair use' in the United States, copyright exceptions in the European Union and provisions regarding text and data mining in the European Union, as well as whether news media organizations should be compensated for platform use of 'snippets' and other texts, are not examined in-depth in this report, but see Section 2, Chapter 2 and Section 4.5, Chapter 6 for a discussion on compensation.

[104] Schaake & Fukuyama (2023); see Gillespie *et al.* (2023, p. 4), for an expanded research agenda on content moderation, arguing for grasping 'the breadth and depth of moderation, across the entire ecosystem of content provision and deep into the infrastructural stack of distribution'.

[105] Gillespie (2020).

[106] Innerarity (2024, p. 10).

It is essential to puncture the '"fallacy of AI neutrality" – represented by the mistaken belief that AI systems can be designed in an inherently unbiased and neutral manner'.[107] Research shows that content moderation and curation systems suffer from biases and encode non-transparent decision-making processes. They are optimized for engagement, that is, to personalize and distribute content to audiences that the system predicts will engage them meaningfully.[108] This means that they can be designed and deployed to achieve political ends, and in ways that exacerbate individual and societal risks. GenAI content is distributed in this way, even though a substantial number of platforms do not have sufficiently detailed policies in place, and may not adhere to them when they do.[109]

There are concerns about the political sensitivity of LLMs and their potential to deepen societal divisions.[110] These tools are relatively new and are being updated quickly, but research shows that the output of three LLM-based chatbots (ChatGPT, Bing Chat and Bard) seems to exhibit varying degrees of bias in response to political queries concerning authoritarian regimes. This is influenced by the language of the prompt. Significant disparities have been found regarding chatbot answers, with Russian language queries resulting in evasive answers regarding content that can be viewed as critical of Russian authorities. Anecdotal evidence shows that this applies to similar queries in Mandarin Chinese on issues such as the persecution of Uyghurs.[111]

Human moderators still play a role in rechecking certain automated decisions and, depending on the jurisdiction a social media company operates in, become active once a user requests that a content-related decision is reviewed.[112] These 'cognitive assemblages' involved in content moderation have been described as a 'cobbled space of pre-emptive calculation'.[113] In all, the trend clearly goes towards more automated moderation, especially in areas where the law is regarded as being clear, such as terrorism content.[114] Where the law is less clear, as in the case of mis- and disinformation, automated tools focus less on content and more on markers related to the distribution channel or the behavior of the account from which the content was launched.[115] Researchers criticize that even 'well-optimized' moderation and curation systems can 'exacerbate, rather than relieve, many existing problems with content policy' because they increase opacity, complicate 'issues of fairness and justice in large-scale sociotechnical systems and … re-obscure the fundamentally political nature of speech decisions being executed at scale'.[116] As discussed, since content produced by AI systems can exhibit and/or reinforce biases against historically marginalized and minority groups,[117] safeguards need to be implemented to prevent these systems from intensifying existing societal inequalities, along with efforts made to use these systems to help elevate the representation of underrepresented groups in the content produced.[118] Efforts to promote ethical standards and diversity in development teams are part of the solution but are not themselves sufficient.[119]

The 'hyper-personalization' of content curation systems attracts much criticism in the literature. Some researchers fear that they may lead especially vulnerable media consumers, such as children and young adults, into 'rabbit holes' of potentially harmful content, among many other harms.[120]

[107] Verhulst (2023, p. 1).
[108] Sančanin & Penjišević (2022).
[109] Issues of the weaponization of information are discussed in Section 4.3.3, Chapter 2, and impacts of content moderation practices are discussed in Section 2.3, Chapter 7.
[110] Biju & Gayathri (2023).
[111] Urman & Makhortykh (2024).
[112] The role and effectiveness of human oversight is discussed in Section 2.1, Chapter 7.
[113] Crosset & Dupont (2022, p. 10), supported by the Fondation du Risque (Allianz, Axa, Groupama and Société Générale) in partnership with the Institut Mines-Télécom and Sciences Po.
[114] Haas & Kettemann (2024); Macdonald *et al.* (2019).
[115] Bontridder & Poullet (2021).
[116] Gorwa *et al.* (2020, p. 1), supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada.
[117] Ross Arguedas & Simon (2023).
[118] Forum on Information and Democracy (2024a).
[119] The weaknesses of these efforts are discussed in Chapter 8.
[120] Amnesty International & AI Forensics (2023). These issues are discussed in Section 4, Chapter 5.

'Harmful' is a difficult criterion to use as a basis for assessing platform content policies. For example, there are few globally accepted examples of prohibited speech. Much 'hate speech', for instance, falls under the protection of free speech in rules in some jurisdictions such as the United States and, depending on the jurisdiction, there are different definitions of illegal speech.[121] This is why it is sometimes argued that automated systems would work better if there was global consensus or a largely agreed on definition of what to find or filter, as in certain cases of terrorism and terrorism financing.[122] Any such effort to forge consensus is likely to be disputed due to cultural and political differences and, even if achieved in the framework of international human rights obligations, may not be translated consistently into practice.

AI systems have also been used, for example, to improve crisis communication.[123] A study by the Organization of Security and Cooperation in Europe (OSCE) suggested that states should mandate platforms to undertake 'crisis-sensitive human rights due diligence', 'crisis-sensitive human rights risks and impact assessments' and emergency measures. Any platform action should 'consider proportionality and reliability on AI tools and automated measures'.[124] Globally, crisis-sensitive human rights approaches by private actors have been urgently demanded.[125] Similar obligations are outlined in Europe's new digital rules, such as the Digital Services Act of 2022, which contains obligations for platforms to conduct risk assessments as to the impacts of their rules and moderation practices on values, including societal cohesion, public health and democratic decision-making processes.[126]

Given that platforms use AI systems for content governance, it is best practice (and legally required in certain jurisdictions, such as the European Union) that they should inform users. However, research shows that users tend to trust moderation decisions less when they know they are automated.[127] This showcases the complexity of achieving the responsible visibility of automated content governance, and user trust is also conditioned by education background and the sociopolitical setting.[128]

One approach is to increase meaningful oversight, including external control over algorithmic systems.[129] This intervention into the private communication realm by platforms, governed by terms of service and algorithmic systems, can be legitimized by reference to the increasing impact of these norms and practices on public values that need to be integrated into the systems. Expert panels or selected user groups, sometimes referred to as platform councils or social media councils, have been suggested. Meta's Oversight Board is one of the early efforts to make the governance of a commercial platform more inclusive of external input.[130] The impact on Meta itself tends to be judged as largely positive, if not very effective, and the Board has been described as overseeing 'one of the largest speech systems in history'.[131] However, the Board has not had substantial cross-industry influence, and has been unable to substantially change the speech governance priorities that Meta exhibits.[132]

3.3 AI SYSTEMS AND NEWS MEDIA

Content created by GenAI can benefit news media diversity by contributing to the efficiency of content generation in specific contexts, and by

[121] Gillespie (2020); see also Galli *et al.* (2023).
[122] Haas & Kettemann (2024).
[123] On the substantial field of research on crisis communication, including communication strategies using social media, see Coombs & Holladay (2022); Jin & Austin (2022). The impacts of social media on conflict escalation are discussed in Section 4.3.3, Chapter 2, on the weaponization of information.
[124] Haas & Kettemann (2024, p. 9).
[125] Fatafta (2024).
[126] This legislation is discussed in Chapters 6 and 7.
[127] Ozanne *et al.* (2022), funded by the US Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) project. See also Chapters 5 and 7.
[128] Kim & Moon (2021).
[129] Nahmias & Perel (2021). Impacts on polarization are discussed in Section 4.4, Chapter 2. Various forms of oversight, including fact-checking, are discussed in Chapter 7.
[130] Kettemann & Schulz (2023).
[131] Douek (2024, p. 373).
[132] Ang & Haristya (2024); Douek (2024); Gulati (2023). Boards such as Meta's examine specific cases of content moderation judgments. Broader forms of oversight aimed at increasing accountability are limited by researcher access to relevant data is discussed in Section 3.5, Chapter 9.

supporting journalists in optimizing the circulation of content contributions after publication, since articles can be published in cross-media formats without substantial additional costs. Research points to the potential of GenAI to 'synthesize broadcast videos using news text during a news broadcast' with better results than manual generation.[133] However, the implementation of these systems requires time and investment, and gains in efficiency and productivity should not be assumed.[134] As the use of GenAI becomes widespread, this can alleviate the burden of relying on overworked newsrooms by automating certain, more mundane, reporting tasks. However, the challenges of news organizations' use of these systems need to be addressed if trust in news media output is to increase, and these organizations are to adhere to ethical standards of data collection and the principle of universality, in contrast to promoting personalized news and other content.[135] AI systems also have a bearing on freedom of expression when they influence editorial decisions, especially when there is a conflict between the editorial need for autonomy and goals that AI tools are optimized for.[136]

The adoption of AI tools by news media organizations for content creation is a concern due to the growing dependency of news media organizations on these technologies.[137]

A survey published in 2023 indicates that GenAI tools, such as ChatGPT, were being used in 49% of newsrooms worldwide.[138] Countries in the Global North and China are leading innovation in AI newsrooms, and research on adoption mainly focuses on the Global North.[139]

Once content is created, news organizations are increasingly dependent on the AI systems used by digital platforms for distribution or circulation. This dependence raises the need for attracting audience traffic that is stimulated by algorithmic personalization. The effects of the interaction between audience traffic and the means to increase the flow of this traffic have implications for the production and visibility of content.

The question of who or what curates content online takes some of the power away from the hands of journalists, the traditional gatekeepers. Platform selection mechanisms usually involve a combination of algorithmic curation (based on criteria specified by business managers) and human editors, making it unclear what the core values underlying selection decisions are, and to what extent they reflect core democratic principles.[140] The impact of curation systems is especially sensitive in public service media (PSM) environments that have a mandate to reach a broad public.[141] Research conducted in France, Germany, Greece, Italy, Poland and Sweden emphasizes that the news media's growing dependence on algorithms means that those who access news media online to meet their information needs do so despite their concerns about the risk of encountering mis- and disinformation.[142] Certain platforms have started to deprioritize news and favor more personal or emotionalizing content.[143] Weaker distribution of accurate information is associated in some studies with more polarized and polarizing media consumption behavior.[144] Platform algorithms using AI tools play a big role in shaping news distribution.[145] It is clear that some news organizations depend heavily on online traffic driven by third-party digital services, leading to dependency on social media for news distribution,

[133] Wu *et al.* (2023).
[134] Simon (2024); Simon & Isaza-Ibarra (2023).
[135] Horowitz *et al.* (2022); Ross Arguedas & Simon (2023); Vaccari & Chadwick (2020). Issues of changes in journalism practices are discussed in Section 4.1, Chapter 2 and of news media content moderation in Section 3.2, Chapter 7.
[136] Helberger *et al.* (2020).
[137] Simon (2022); see also the survey of AI guidelines for media across 17 countries in de Lima Santos *et al.* (2024), supported in part by the European Commission.
[138] WAN-IFRA (2023).
[139] Beckett & Yaseen (2023); and see Beckett (2019); Kothari & Cruikshank (2022); Marconi (2020).
[140] van Dijck *et al.* (2018b).
[141] Horowitz *et al.* (2022).
[142] Schaetz *et al.* (2023), supported by the Federal Ministry of Education and Research (BMBF, Bundesministerium für Bildung und Forschung), Germany.
[143] Meese & Hurcombe (2021).
[144] Schirch (2021); see also the discussion on polarization in Section 4.4, Chapter 2.
[145] Meese & Hurcombe (2021); van Dijck & Poell (2013).

a trend that does not uniformly affect the entire industry.[146]

Facebook/Meta's approach to news. Facebook's interest in news content has grown as it sought to monetize online advertising and counter X (then Twitter)'s emerging status as a key news source. In 2013 the company began promoting news publishers' content in its personalization system. This encouraged news organizations to focus on Facebook distribution strategies for their news. Facebook developed technologies for hosting content directly (e.g., the launch of Instant Articles), and incentivized publishers to keep their content on its platform. The platform's shift to video content and the introduction of Facebook Live led the media industry to adapt to these changes. The relationship between publishers and Facebook soured due to monetization challenges, inflated video metrics by Facebook, and controversies surrounding mis- and disinformation, especially during the 2016 United States presidential election. Facebook's response was to step away from news distribution in 2018, changing its News Feed algorithm to prioritize personal content. Faced with this challenge, some news media organizations altered their distribution strategies, aiming to regain control of revenue streams and favor core audience interests over Facebook demands.[147]

The extent to which the push to adopt AI tools will increase news media dependency on digital platforms is unclear.[148] Claims that the 'AI goldrush' will increase the potential for infrastructure capture and shift even more control to platform companies raises questions about control, dependence and autonomy, as the adoption of AI tools in newsrooms extends platform control over the news production processes and the distribution networks.[149] While there is a growing market for AI tools to cater to newsrooms' needs, with smaller players such as Narrativa, Retresco, Adobe and others trying to position themselves in the market, the dominant players operate in an oligopolistic market (see Table 3.1).

Table 3.1
AI systems uses in the news media gatekeeping process

Production and distribution process	Use of AI systems
Access and observation	• Information discovery. • Audience and trends analytics; story detection. • Prompting for new ideas following from a news story.
Selection and filtering	• Verification, claim matching, and similarity analysis (e.g., for fact-checking). • Content and/or document categorization; analysis of datasets. • Automated collection and analysis of structured data (e.g., financial, banking, and sports data). • Coding assistance for various tasks. • Transcription and translation of audio and video. • Search in archives and/or metadata.
Processing and editing	• Brainstorming and ideation. • Content production (writing of draft text or articles; editing of news content). • (Re-)formatting of content for online, social media, print, broadcast (e.g., summarization, simplification, stylistic changes; text-to-video, speech-to-text, text-tospeech translation). • Copy editing, adaptation to house style. • Tagging of content, headline, and SEO [search engine optimization] suggestions.
Publishing and distribution	• Personalization and recommendation. • Dynamic paywalls, audience analytics. • Content moderation.

Source: Simon (2024, p. 13).

Multiple factors influence the extent of news organizations' dependence on digital platforms and their AI tools. These include the country (there is, for example, weak evidence of dependence in Germany); the kind of news organization; whether organizations are established, legacy or digital only (a study in South Korea, for example, found that legacy organizations experienced greater pressure than digital only); and how PSM addresses its public role and its relationship to audience reach (e.g., to young people).[150]

[146] Bakke & Barland (2022).
[147] Meese & Hurcombe (2021).
[148] Simon (2022).
[149] Simon (2022).
[150] See (Hase *et al.*, 2023) whose findings are challenged by Eichler (2023); see also Poell *et al.* (2023); Pyo (2022); van Dijck *et al.* (2018a).

The wave of enthusiasm surrounding AI systems centers around its potential to transform the social roles of journalism, especially as it supports the profession's core functions in a democracy.[151] Analysis of media coverage of 'AI' use in journalism over a five-year period in the United Kingdom and the United States indicates that opinions are far from uniform.[152] There is a tension between the industry (newsroom leaders and funders) advocating for the use of this technology long-term, and professionals (journalists) highlighting concerns, for example, about the impact of AI systems on accuracy, fairness and transparency.[153] The use of AI systems in journalism is normatively evaluated in relation to stages of news work – information gathering, selection and production, and distribution and consumption, normative dimensions of accuracy, accessibility, diversity, relevance and timeliness.[154]

Contributions of AI systems to fulfilling journalism's democratic role. Discussions on how to develop AI tools responsibly should be grounded in a normative perspective on the underlying values and principles, including the need to start with identification of values and principles with multiple stakeholders; the development of a forward-looking vision on the role of journalistic AI, grounded in a normative framework focused on editorial mission, fundamental rights and the democratic role of the media; and understanding how journalists, editors, managers, developers, users and other stakeholders can be empowered to become active agents in decision-making processes around the implementation of journalistic AI.

The argument for a more inclusive decision-making process comes from the realization that AI apps are not just tools, but integral components of the public communication infrastructure, whose design is of concern to all stakeholders. The challenge is 'to design decision-making routines so that they become more accountable to the public, more inclusive and cognizant of diverse and underrepresented voices in society, and less dependent on a small number of major technology companies'.[155]

Ethical concerns underlying the adoption of AI systems by journalists include whether automated content is consistent with editorial criteria; personalization that respects diversity and promotes a thriving public sphere; monitoring the quality of data to avoid bias; responsible safeguarding of user privacy; quality journalism with an emphasis on the human factor; funding of platforms and journalism independence; and AI systems to foster the values of journalism.[156] The Council of Europe's Steering Committee on Media and Information Society has published its *Guidelines for the Responsible Use of Artificial Intelligence in Journalism*, and there are numerous codes of practice to guide the use of the these technologies.[157] There are concerns that the inclusion of AI tools in journalism routines could shift moral and editorial responsibility away from newsrooms, with consequences for public perceptions of news media bias.[158] A study of professionals in newsrooms in 16 countries in the Asia Pacific, Europe, Latin America, the Middle East and North Africa (MENA), North America and sub-Saharan Africa regions found that ethical concerns were significant. More than 60% of respondents were concerned about editorial quality, and many expressed a desire for AI systems transparency and the implementation of ethical guidelines.[159]

[151] Lin & Lewis (2022), supported by the Ministry of Science and Technology, Taiwan; Moran & Shaikh (2022); Beckett & Yaseen (2023), supported in part by Google News Initiative.
[152] Moran & Shaikh (2022).
[153] Beckett & Yaseen (2023).
[154] Lin & Lewis (2022), supported by the Ministry of Science and Technology, Taiwan.
[155] See Helberger *et al.* (2022, p. 1621), supported by the European Research Council (ERC).
[156] Pocino (2021, p. 19).
[157] Council of Europe (2023).
[158] Calice *et al.* (2023); Moran & Shaikh (2022).
[159] Beckett & Yaseen (2023), supported in part by Google News Initiative.

In addition, developing AI model de-biasing techniques has been found to be very challenging for journalists in other studies, and the capacity to address this issue depends on the data quality that is available to journalists in their work.[160]

3.4 USE OF GENERATIVE AI BY MIS- AND DISINFORMATION ACTORS

The widespread adoption of AI systems for content generation and distribution is associated with an increase in the spread of mis- and disinformation.[161] In their response to the draft amendments to the IT rules in 2021, in India, IT for Change emphasized that 'approaches to addressing misinformation and fake news need to be reframed with due cognizance of the information economy and its technological mechanics'.[162] The accessibility and sophistication of content produced by GenAI are increasing as these tools provide creative possibilities for producing or altering textual, visual, auditory and audiovisual data, and are used by both private and state actors.[163]

A survey conducted by Freedom House in 2023 found that a minimum of 47 countries employed commentators to manipulate online discussions in their favor, which is double the number of countries involved a decade ago.[164] As indicated, the evidence on how systematic these efforts are and which specific actors are involved is missing or weak, despite the fact that these 'disinformation tactics' are growing in sophistication as GenAI tools become more powerful, readily accessible and user-friendly. It is clear that they are being used to foment uncertainty, defame adversaries and sway public discourse.

Figure 3.2
Example of realistic AI-generated face using the 2020 algorithm StyleGAN2

Source: Authors of report.

The proliferation of false information, propaganda and hoaxes has grown dramatically with the spread of the internet and social media. It increased further with the use of user-friendly, GenAI tools, enabling 'deepfake' creators to build realistic synthetic videos, audios or images of real individuals without extensive technical expertise or substantial financial resources. For example, CounterCloud – an AI model said to produce automated disinformation that is convincing 90% of the time – is reported to be usable at a cost of less than USD 400 per month.[165] This illustrates the cost-effectiveness and simplicity with which significant mis- and disinformation operations can be generated (see Figure 3.2).

In the United States, AI-generated information has been used to tarnish the reputations of political rivals. In Venezuela, state-controlled media used AI-generated videos featuring fabricated news anchors from a fictitious international English language network to disseminate pro-government

[160] Dierickx *et al.* (2023b).
[161] See the reports under the EU Code of Practice on Disinformation, March 2024, at https://disinfocode.eu.
[162] Rajkumar & Ashraf (2023, p. 3), IT for Change is an independent NGO, Bengaluru, India.
[163] Bontridder & Poullet (2021).
[164] Funk, Shahbaz & Vesteinsson (2023) supported by Amazon, the Dutch Ministry of Foreign Affairs, Dutch Postcode Lottery, Google, the Hurford Foundation, the Internet Society, Lilly Endowment Inc., the New York Community Trust, the US State Department Bureau of Democracy, Human Rights and Labor (DRL) and Verizon.
[165] Funk, Shahbaz & Vesteinsson (2023), supported as above.

propaganda. Produced by Synthesia, a company specializing in the creation of personalized deepfakes, this content was widely shared on social media platforms. In 2023, during the Nigerian elections, a modified audio recording created using GenAI was shared on social media. The recording falsely claimed to provide evidence of an opposition presidential candidate's involvement in attempts to manipulate the ballots.[166]

Ofcom's Online Nation report in 2023 found that two-thirds of online 16- to 24-year-olds and over half of 25- to 34-year-olds in the United Kingdom were worried about the future impact of GenAI on society,[167] reflecting a new phase in the public's growing distrust in digital technologies.[168] A report by the United Nations General Assembly concluded that AI-generated mis- and disinformation could 'undermine information integrity and access to information' and 'undercut the protection, promotion and enjoyment of human rights and fundamental freedoms'.[169]

Some researchers argue that concern about the risks of AI-enabled mis- and disinformation is exaggerated, and that it distracts attention from other issues. In one study, the authors note evidence that heavy misinformation consumption is limited to people who are already more likely to seek it out, leading them to conclude that increased information quality is unlikely to have a significant effect (see Figure 3.3).[170] We should be wary, however, of assuming that such conclusions apply globally.[171] A study in sub-Saharan Africa found that people displayed a greater willingness to share mis- and disinformation compared to those in the United States.[172]

The potential impact of GenAI on mis- and disinformation can occur in four categories: (1) increased quantity; (2) increased perceived quality; (3) increased personalization; and (4) accidental generation of plausible but false information.[173] Measuring the scale of AI generation and the distribution of mis- and disinformation and the impact of mis- and disinformation campaigns is challenging because of the difficulties of identifying, gathering and analyzing data that fully reflect people's day-to-day online experiences. The evidence that does exist suggests that the scale of AI generation and distribution of mis- and disinformation grew significantly in the five years to 2023.[174] Past empirical studies of bots, for example, have concluded that they are 'omnipresent' on social media platforms such as X (formerly Twitter),[175] although many are used for relatively benign purposes.[176] A study in 2019 identified 'cyber troop' (government or political party actors tasked with manipulating public opinion online) activity in 81 countries.[177]

[166] Repucci & Slipowitz (2022) supported by Google Inc., the Hurford Foundation, Jyllands-Posten Foundations, Lilly Endowment Incl, Meta Platforms Inc., and National Endowment for Democracy; Ryan-Mosley (2023).
[167] Ofcom (2023d).
[168] Dutta & Lanvin (2023).
[169] UN (2024c, p. 3); see also UN (2024b). This is discussed in greater depth in Chapter 5.
[170] Broniatowski *et al.* (2023); one author is from the Office of the Assistant Secretary for Financial Resources (ASFR), US Department of Health and Human Services; see also Motta *et al.* (2024).
[171] Madrid-Morales & Wasserman (2022).
[172] Wasserman & Madrid-Morales (2019), supported by the National Research Foundation (NRF), South Africa.
[173] Simon *et al.* (2023).
[174] Funk, Shahbaz & Vesteinsson (2023).
[175] Keller & Klinger (2019).
[176] Makhortykh *et al.* (2022).
[177] Bradshaw & Howard (2019), supported by the European Research Council (ERC), Hewlett Foundation, Luminate and Adessium Foundation.
[178] As discussed in Section 4, Chapter 2 and Section 2, Chapter 5.

Figure 3.3
Deepfake image of Donald Trump generated using Stable Diffusions

Source: Authors of report.

The scale of the mis- and disinformation that is generated and amplified as a result of the use of AI systems is difficult to measure, and there is a lack of consensus as to its impact, and relatively limited evidence on its impact on trust in information and news media news.[178] Sources of evidence are variable in quality, level of detail and overall reliability. They include incidence reports, some corporate case studies, some surveys of worldwide campaigns, such as the Global Inventory of Organized Social Media Manipulation, and reports on content takedowns by platforms such as Facebook.[179] A lack of standards for, and transparency in, data collection, makes it difficult to verify and replicate findings.[180] The United Nations Policy Brief on information integrity on digital platforms documents several cases of harm linked to mis- and disinformation, including violence against individuals and groups ensuing from the posting of hate speech, reduced take-up of Covid-19 vaccination programs and risks for the achievement of the Sustainable Development Goals.[181] A report on a 2022 survey found that 70% of United Nations peacekeepers felt that mis- and disinformation was having a 'severe, critical or moderate impact on their work'.[182] A synthesis of evidence from 1,300 sources (news articles, academic papers, white papers, and a range of other grey literature) found case studies of impact in over 70 countries.[183] Studies of the proliferation of mis- and disinformation during the Covid-19 pandemic concluded that there was a significant impact on vaccine uptake.[184] However, attempts to develop models to simulate the potential impact of mis- and disinformation face challenges even when data access becomes easier, in part, because there are substantial issues to be overcome in modeling real data, and many events in the world can affect how exchanges take place on platforms.[185]

3.5 COUNTERING MIS- AND DISINFORMATION

The absence of robust AI content classification has enabled both state and for-profit actors to exploit the tendency of personalization systems to prioritize engagement-rich content.[186] Authoritarian nations are using AI systems to broaden and reinforce censorship. Research by Freedom House identified 22 countries that have enacted legislation mandating or providing incentives for internet platforms to use AI to eliminate speech on the internet the state deems undesirable;[187] for example, chatbots in China are programmed not to react to inquiries about Tiananmen Square. YouTube and X were required by the Indian government to restrict access to a documentary that showed the violence that occurred when Prime Minister Modi

[179] See Windwehr & York (2020). Facebook publishes annual transparency reports documenting its content moderation actions. These have been criticized for not disaggregating the types or including precise quantity of content removed. See Bradshaw *et al.* (2021), supported by the European Commission, European Research Council (ERC) and the Adessium Foundation, Civitates Initiative, Ford Foundation, Hewlett Foundation, Luminate, Newmark Philanthropies and Open Society Foundations. See Bradshaw *et al.* (2020) for country case studies and a global inventory of organized social media manipulation.

[180] Kostygina *et al.* (2023), supported by the National Cancer Institute and National Institute on Drug Abuse of the National Institutes of Health (NIH), US.

[181] UN (2023a).

[182] UN (2023a, p. 13).

[183] Bradshaw *et al.* (2020, 2021), supported by the European Commission, European Research Council (ERC) and the Ford Foundation.

[184] Fertmann & Kettemann (2021); Naeem *et al.* (2021); Posetti & Bontcheva (2020).

[185] Lamnitchi *et al.* (2023), funded by DARPA (Defense Advanced Research Projects Agency), US.

[186] Thomas (2022).

[187] Funk, Shahbaz & Vesteinsson (2023).

was Gujarat's chief minister. The Indian government has also urged technology companies to employ AI-based moderation techniques to regulate content.[188]

Research has been funded by the European Commission to produce tools identifying mis- and disinformation, and several of these are used by professionals in their fact-checking work, but research also shows that tools on their own cannot counter the threat of mis- and disinformation.[189] As a 2023 study on anti-disinformation responses shows, tackling this information requires a unified effort that transcends individual stakeholders, such as governments acting through laws, and platforms acting through their terms of service.[190]

Governments need to provide a legal framework for removing illegal content, and an accountability and transparency framework for problematic content, internal rules and algorithmic personalization systems, and these need to be enforced. Governments must also secure adequate funding for researchers and civil society to leverage data access rights. Additionally, promoting partnerships with digital platforms can help elevate verified information sources, supporting PSM and independent entities that contribute to democracy and education.

4 AI Systems and Democracy

This section addresses the reciprocal relationships between the development and deployment of AI systems and mediated public sphere(s), including how these relationships affect news media diversity and media freedom, and more generally, the interaction between these systems and societal resilience and cohesion, social and environmental sustainability.

4.1 AI SYSTEMS AND MEDIATED PUBLIC SPHERE(S)

The use of AI systems for content governance shapes the public sphere(s) in which communication flows occur. While private communication platforms that use these technologies do not themselves directly 'censor', the design and use of content governance algorithms influences democratic discourses.[191] Just as AI systems can contribute to more diverse information ecosystems, they can reinforce the monitoring capabilities of authoritarian states and enhance inequalities and unfair power structures through labor extractivism.[192] Without negating the role of automated tools, it is important to realize that non-technology-related phenomena, such as the quality of a social security system or whether gender equality is supported, are found to be bigger factors when it comes to furthering societal cohesion and resilience. This means that, in assessing the impact of AI systems, the socio-economic and political context in which information ecosystems operate have to be taken into account, as well as the policy and regulatory situation of a country or region.[193]

Notwithstanding these broader considerations, it is important to account for some of the specific influences that AI systems can have on the composition and functioning of the public sphere. AI tools used by platforms to curate content tend to favor emotionalizing content that can be used to increase engagement. This can reward social and political groups that communicate substantially through this content, or in that style.[194] Geographically dispersed and fringe

[188] Ryan-Mosley (2023).
[189] EC (2024b); Teyssou *et al.* (2017); Marinova *et al.* (2020), partially supported by the European Commission.
[190] Berger *et al.* (2023a). Measures including legislation, platform policies, fact-checking initiatives and literacy training aimed at achieving greater control over the creation and spread of mis- and disinformation are discussed in detail in Chapters 5, 6 and 7.
[191] Elkin-Koren (2020), supported by the Israel Science Foundation.
[192] Boix (2022); Adams (2022), prepared by an independent, non-partisan, African think tank.
[193] Breuer (2024), supported by the European Union Horizon 2020 program and Federal Ministry for Economic Cooperation and Development (BMZ, Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung), Germany; see also Birwe (2024).
[194] Noble (2018).

groups can profit from easier connections through social media.[195] Personalization systems tend to amplify content algorithmically that emotionalizes and divides because platform business models demand engaging content.[196] Social media use is positively correlated with more diverse information consumption in some studies,[197] for example, and the use of interest histories (personalization based on previous behavior) to shape information consumption has been found to increase content diversity.[198] However, platforms receive 'outsized attention and criticisms' for being the main drivers of societal polarization, when it is also important to take account of the broader societal conditions.[199] Some argue that underlying societal inequality is a bigger threat to societal polarization, emphasizing that the relationship between AI systems development and societal conditions is reciprocal, but also characterized by power asymmetries.[200]

The protection of democratic values supporting the existence of the public sphere in the face of technological change is a key goal of the regulatory processes around platforms.[201] An increasing emphasis on user rights-related obligations for platforms has emerged since the early 2000s in court rulings and laws, especially in the European Union.[202] Power asymmetries between the Global North and Global Majority World give rise to key areas of conflict that are contributing to an 'AI divide'. These include the increasing use of AI systems in Global Majority World countries, where there is a lack of investment in the underlying information ecosystem infrastructure and in content moderation capacities, for example, for smaller language communities and non-Global North cultures. These conditions are coupled with workforce 'extractivism' – the use of low-wage 'ghost' workers for training AI models.[203]

Political messaging and GenAI. Evidence of the use of GenAI for creating mis- and disinformation in political messaging is growing. This may be due partly to the increasing availability and low cost of GenAI tools whose use requires little or no technical expertise. Some evidence suggests that, while tools for detecting mis- and disinformation can do so with an accuracy of 80-90% on GenAI content created in the Global North, they are much less effective on content created in Global Majority countries because of biases in their training data. According to Sam Gregory, program director of the non-profit organization WITNESS: 'As tools were developed, they were prioritized for particular markets', and the data used to train the models, 'prioritized English language – US-accented English – or faces predominant in the Western world'.[204]

Any discussion on the democratic implications of AI systems needs to include Global Majority World voices, and develop alternatives to current practices of exercising socio-economic and geopolitical power through algorithmic tools and datafication.[205]

Calling only for transparency of content governance systems that influence global information distribution processes *is not enough*. Engagement with diverging approaches to making algorithms used by communication actors more accountable is necessary.[206] Key transparency challenges include information asymmetry, uncertainty and resourcing. This requires interdisciplinary engagement and decisions about legal rights to access information,

[195] Kreiss & McGregor (2023).
[196] Bail (2021); Settle (2018).
[197] Gil de Zúñiga *et al.* (2021); Möller *et al.* (2018).
[198] Möller *et al.* (2018).
[199] Kreiss & McGregor (2023).
[200] Kreiss & McGregor (2023).
[201] Mökander *et al.* (2023), supported by AstraZeneca. Platform regulation is discussed in Section 4.3, Chapter 6.
[202] Katzenbach (2021), funded by the European Commission.
[203] Monasterio Astobiza *et al.* (2022).
[204] Elliott (2024).
[205] Roche *et al.* (2023), funded by the Science Foundation Ireland (SFI), Centre for Research Training in Artificial Intelligence; Ricaurte (2022). See also Chapters 4 and 8.
[206] Ananny & Crawford (2018).

shared decision-making about AI transparency choices, efforts to understand social and societal impacts, and adequate resourcing of transparency teams and audits.[207] It is uncertain the extent to which the voices of those in Global Majority World regions will play a role as countries in the Global North push back on the United Nations' efforts to give countries in these regions – including China – a strong voice in AI systems governance.[208]

In the light of the challenges of algorithmic content production and distribution, media plurality and media diversity, as well as media freedom, must be protected. The existence, and plurality, of independent news media of sufficient quality is impacted by increased use of AI tools for content production which, in turn, is influenced by trends in market concentration triggered by AI systems investment.[209] In 2018 the Council of Europe recommended that automated decision-making processes governing the distribution of online content should 'improve the effective exposure of users to the broadest possible diversity of media content online'.[210] Assessing how to measure media diversity is not a simple task, and proposals for metrics aimed at assessing initiatives to support a more diverse media environment are only a first step.[211]

Research that suggests AI systems use in social media has negative effects on content diversity in terms of its distribution may neglect the multidimensionality of diversity that encompasses 'topic plurality, genre' and 'plurality in tone'. Studies using the concept of 'exposure diversity' – the diversity of information users actually see – find that algorithmic personalization systems have strong positive effects on diversity. The 'element of surprise: serendipity' is an essential part of (most) of these systems. Highly personalized systems that increase the perceived relevance of specific content for users can reduce the range of information they encounter, although increases in media and information (and AI) literacy may mitigate this effect.[212] Regulatory approaches, discussed in Chapter 6, aim to address the need to receive data from platforms on key optimization goals of content governance systems.

4.2 AI SYSTEMS AND SOCIETAL RESILIENCE AND COHESION

Information ecosystems are connected to other societal systems, and although AI systems are only one factor in societal transformation processes, they can both challenge and enhance societal resilience and cohesion. Societal resilience refers to the ability of a society to react to, and recover from, challenges and disruptions, including short-term disruptions (e.g., armed attacks), medium-term crises (e.g., the Covid-19 pandemic) and long-term challenges (e.g., climate change).[213]

Societal cohesion is a key contributing factor to, and predictor of, societal resilience. It refers to the capacity and extent to which a society's members cooperate and work together toward collective well-being based on shared values. Values are shared, questioned and developed through communication processes. When automated content moderation tools play an important role in information ecosystems, they can have an impact on societal cohesion and thus resilience. Being aware of the rules and practices governing mediated discourse is important for meaningful democratic participation, and increasing 'algorithmic awareness' is an important aspect of AI literacy.[214] By increasing sensitivity to the impact of AI systems on content production and distribution, societal cohesion and resilience can be better supported. Conversely, research suggests that greater societal resilience is positively correlated with resistance to mis- and disinformation.[215]

[207] Ruffo *et al.* (2023), funded by IBERIFIER (Iberian Digital Media Research and Fact-Checking Hub), European Digital Media Observatory (EDMO); Bates *et al.* (2023).
[208] Alexander (2024) argues that the United Nations only ostensibly seeks to give Global Majority World actors a louder voice.
[209] See Section 2, Chapter 2 for structural conditions affecting the financial sustainability of news media and their dependence on platforms that deploy AI tools.
[210] Council of Europe (2018, para. 2.5); see also Heitz *et al.* (2021).
[211] Ranaivoson *et al.* (2022).
[212] Helberger *et al.* (2018), supported by the European Research Council (ERC); Möller *et al.* (2018), supported by the European Research Council (ERC); Kreps *et al.* (2022). See Chapter 5 for a discussion of literacy.
[213] Berger *et al.* (2023b); Haas & Kettemann (2024); Kettemann & Lachmayer (2021); Veale *et al.* (2023).
[214] De Vivo (2023). AI literacy is discussed in detail in Chapter 5.
[215] Kertysova (2018).

Evidence also suggests that use of AI by political actors can increase the quality and speed of responses to political queries by citizens, thus leading to a higher level of engagement, with the important caveat that citizens must be helped to understand and trust how these systems are used.[216]

There is contradictory evidence concerning whether automated content governance is a main driver of societal polarization, and hence a decline in social cohesion, although polarization dynamics is a key field of research.[217] A lack of digital literacy, societal vulnerability towards 'information pollution', and a preexisting 'fragmentation' of society are cited as playing more substantial roles. However, the prevalence of mis- and disinformation such as hate speech can lead vulnerable groups to withdraw from online discourses, thus decreasing social cohesion.[218] There is little evidence that automated content governance systems are the only contributor to polarization, but news diversity and media consumption practices can clearly be affected by 'machine gatekeeping'.[219] Exposure to misinformation and partisan information also can elicit strong emotions, which, in some studies, is shown to lead to some 'attitude polarization', as discussed in Chapter 2.[220]

4.3 AI SYSTEMS AND SOCIAL SUSTAINABILITY

The integration of AI systems into the workplace is profoundly transforming labor conditions across industries. Content moderation is essential for maintaining the quality and safety of online platforms but, despite its importance, the job is often outsourced to contract workers who face unstable employment conditions.[221] Processes underpinning data collection, content labeling and training can contribute to harm, including traumatization as a result of working with problematic content or training data, which affects underpaid workers in countries that lack stringent labor protection laws.[222]

The Amazon Just Walk Out or Amazon Go stores, where people could 'enter at gate, shop and walk out', were employing hundreds of workers in India, and this prompted the company to roll back the use of this technology in its stores.[223] Uber, Lyft and DoorDash use AI systems and data analytics extensively to manage their operations.[224] Drivers and delivery personnel working for these companies are typically classified as independent contractors rather than as employees. This means that many workers do not receive the benefits or protections associated with employment, such as health insurance, paid leave or job security. The intersection of surveillance capabilities, worker monitoring and labor conditions means that these companies' uses of AI systems and their approaches to collecting and processing data are attracting attention due to the potential impacts on worker privacy, autonomy and rights, with data privacy concerns being raised in certain regions, including Africa.[225]

AI systems managing operations. In India, the door-step food delivery platform Swiggy has gamified insurance for its rider partners. Swiggy's weekly ranking system allows workers to access health insurance depending on the number of 'perfect' deliveries they make. In 2021, Amazon designed a 30-day 'Delivery Premier League' (DPL) for its part-time workers, under the Amazon Flex program. Modeled after the flagship cricket event Indian Premier League, DPL gamifies delivering

[216] Muñoz (2023).
[217] Ruffo *et al.* (2023). See also Section 4.4, Chapter 2.
[218] Nordic Council of Ministers Secretariat (2023); Sîrbu *et al.* (2019); Washington (2023); see also Breuer (2024), supported by the European Union Horizon program and Federal Ministry for Economic Cooperation and Development (BMZ, Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung), Germany.
[219] Evans *et al.* (2023); Ross Arguedas *et al.* (2022a).
[220] Weismueller *et al.* (2023).
[221] Ahmad & Greb (2022).
[222] Veale *et al.* (2023).
[223] AWS (n.d.).
[224] Bitter (2024); Burrell (2016).
[225] Abdulrauf & Dube (2024).

packages. Each hour spent on the platform collecting packages from mini-warehouses and delivering them to customers' homes constitutes a 'run' – a unit of scoring in cricket. The more hours spent delivering, the more deliveries riders accumulate, ultimately resulting in rewards such as smartphones, motorbikes, televisions and Amazon gift cards in addition to the flat 125 rupees (about USD 1.50) paid per hour.

In many Indian cities, local governments have implemented GPS-based systems to monitor sanitation workers to boost productivity and manage schedules, raising concerns about privacy and the dignity of labor. In cities including Patna and Pune, GPS devices are used to track the movements of sanitation workers. Amazon uses sophisticated systems to track the movements and productivity of warehouse workers. Workers are often required to pack hundreds of boxes per hour, and any time spent 'off-task' can lead to warnings or job termination.[226]

4.4 AI SYSTEMS AND ENVIRONMENTAL SUSTAINABILITY

Training state-of-the-art AI models is an energy-intensive process. LLMs demand vast amounts of data and power-intensive training processes, involving complex calculations, run on thousands of high-powered graphics processing units (GPUs) over several weeks. This can lead to a sizable environmental footprint, because data centers are one of the major drivers of increases in energy demand and in greenhouse gas emissions.[227]

Data center energy demand escalates. The rapid development and adoption of AI systems is leading to escalating demands on the digital infrastructure – data centers – that are essential to its progress. Goldman Sachs has predicted that by 2030, data center power energy demand will grow by 160%.[228] An investigation in late 2024 suggested that the real emissions from data centers can be more than six times the officially reported values.[229]

Advances in chip technology can mitigate environmental impacts, offering greater computational outputs per watt of power consumed, thus a potential offset – although relatively minor –to the energy-intensive nature of extensive data operations.[230] Google has used AI systems to enhance the energy efficiency of its data centers, reducing its cooling energy requirements by up to 40%.[231] In some countries resistance to the energy consumption of large data centers and computing resources is emerging and strengthening.[232] This ties in with the general growing demand for public participation in decisions impacting on sustainability agendas. Researchers are calling for holistic approaches to these issues, encompassing the whole lifecycle of AI systems development, including environmentally responsible innovation.[233]

[226] Bansal (2024); Bitter (2024); Christopher (2021); Nagaraj (2020).
[227] iea50 (2024).
[228] Goldman Sachs (2024).
[229] O'Brien (2024).
[230] Berthelot *et al.* (2024); Cowls *et al.* (2023), supported in part by the Vodafone Institute; one author is on the Board of Directors for Noovle S.p.A., Italy.
[231] Burgess (2016); Google (2022).
[232] Velkova (2024); see also WEF (2024).
[233] Brevini (2021); Makan (2023); Wu *et al.* (2022).

5 Chapter Summary

The central question addressed in this chapter is how AI systems development and use is co-evolving with the protection of internationally protected human rights and fundamental freedoms.

The term 'AI' entered popular discourse to describe – misleadingly – a class of digital systems that use AI technologies to perform tasks that were the preserve of human expertise. This chapter has reviewed common definitions of 'AI', and explained why 'machine learning' (ML) is a more appropriate term to describe the systems in use in digital platforms for content governance, and why 'AI' is hard to avoid given the degree to which it has entered common usage. With the advent of GenAI – which can generate new content in the form of text, images and video – the impact of AI systems on people's experiences of information ecosystems as content audiences and consumers is growing.

This chapter has explained how human rights apply in the age of digital transformation and, specifically, how they can be upheld as novel AI systems are developed and applied in different societal fields, ranging from care work to content moderation, from journalism to lending decisions. Although we argue that calls for new human rights are misguided, we emphasize that certain human rights challenges arise specifically through the widespread use of automated content governance and how decisions in this area impact society-wide democratic decision-making processes. The focus in this chapter was particularly on algorithmic bias and fairness, the relationships between freedom of expression, information and the news media, and approaches to privacy protection and participatory rights, all of which are affected by developments in AI systems.

The chapter also looked in some detail at how AI systems are being used for content governance and the impacts of their use on information integrity. Our examination of how AI systems are being deployed for content governance emphasizes that no algorithm or training data set can be free of bias. This has impacts on news media personalization systems, and it also creates new opportunities for the use of GenAI by those who generate and disseminate mis- and disinformation, as well as for the news media industry, with consequences for the public sphere.

Understanding the properties of AI systems, including how these are related to the way they are created, optimized and used, is essential if their impact is to be gauged and if regulation is to be effective.[234] A stronger focus on explainability and accountability best practices for automatic content governance systems is crucial. This is because of the need to achieve greater transparency of AI-enabled decisions through improved understanding and by encouraging trust in AI-enabled decisions when it is warranted. The pace of innovation and adoption of AI systems, and especially the emergence of GenAI, is inevitably creating substantial gaps in knowledge about how these systems are incorporated into information ecosystems and with what consequences for the health of information ecosystems.

The synthesis of research in this chapter shows that:

- It is important for researchers to be specific about the technologies, such as algorithms or ML and LLMs, that are being examined; there is a proliferation of research and commentary that treats 'AI' as a single category, and this is unhelpful in the face of the need to respond differently to the risks of these systems. These vary substantially in terms of the risks they pose for human rights and societal processes of self-determination.

- It is essential to confront rule of law issues, and to take account of the variety of ways in which AI systems become embedded in people's lives, which differ across countries and regions. Discussions about the contribution of AI systems to the health of information ecosystems, or its detrimental effects, need to be as inclusive as possible.

- Internationally agreed and protected human rights and fundamental freedoms are fully

[234] AI systems governance is discussed in Section 4.4, Chapter 6 and Section 3.1, Chapter 7.

applicable in today's information ecosystems, but states need to ensure that their obligations to respect, protect and implement these rights are responsive to the specific challenges posed by the new actors, instruments and power relations in the age of digital transformation.

- Biases in AI systems are a consequence of biases in the (selection of) data on which they are trained. This is not inevitable, but rather the result of human rights-insensitive practices of AI systems developers.

- Focusing mainly on tweaking content governance practices and systems ignores the multifaceted underlying causes of social discord and distrust that give rise to polarized public opinion. A focus on the 'public worthiness' of information, rather than on information 'disorder', is likely to be a more effective way to reveal the complex elements of visibility, access, reflexivity, mediation, influence and information legitimacy.

- There is substantial evidence that the use of AI systems in content governance can lead to rights violations. Content governance systems frame the conditions under which content is seen and with whom it is shared. A lack of transparent training and deployment of automated content governance tools challenges both individual and societal rights, including freedom of expression and information and privacy, as well as democratic participatory rights.

- No single content moderation technique can be acceptable to every online participant and no content moderation or content curation system is neutral or non-discriminatory. These systems are being deployed to (usually) achieve commercial ends, with some social media companies pursuing an additional, sometimes politicized, agenda, or attempting to reduce the prevalence of certain content categories, like political content. Safeguards are needed to prevent the platforms using these systems from intensifying existing societal inequalities, increasing polarization and contributing to information disorder.

- AI systems play an important role in newsrooms in content production and distribution. The personalization of news media may positively influence the diversity of news that online users engage with, but it is essential that algorithms and other AI tools are used transparently and ethically because of their impact on the integrity of information in public sphere. The impacts of these systems on efficiency and productivity in the news industry should not be assumed.

- AI systems are being used by a range of actors to generate and distribute false information, propaganda and hoaxes, but measuring the scale of mis- and disinformation and its impacts remains challenging, partly because of the need to access real data and to develop behavioral models.

- Governments need to provide legal frameworks for defining and removing illegal content as well as assuring accountability and transparency for problematic content, and internal rules and algorithmic personalization systems. The rules arising from these frameworks need to be enforced. The European Union's Digital Services Act and AI Act, and recommendations and conventions from international organizations, including UNESCO and the Council of Europe, offer examples of good practice, but their concrete impact is not yet clear.

- It is essential that research takes account of the reciprocal relationships between the development and deployment of AI systems and the evolution of information ecosystems, including the implications for mediated public sphere(s), societal resilience and cohesion, the social sustainability of labor markets and environmental sustainability.

Research is needed:

- To provide ongoing insight into the way human rights law is being interpreted and applied at the country (regional) level, to assess whether commitments to protect fundamental rights are being met.

- To develop improved understanding of the impacts of decisions throughout the AI

development chain on the health of information ecosystems. The impact of AI systems on how information is spread and amplified by platforms remains poorly understood due to a lack of data, the complexity of interlinked algorithmic personalization systems in use by major digital platforms and diverse country contexts.

- To assess whether improving data diversity, conducting regular algorithmic audits and enforcing transparency is likely to ensure that AI systems are developed and used responsibly and ethically to achieve algorithmic fairness, thus helping to mitigate their potentially harmful effects.

- To undertake detailed studies on the mechanisms of AI-driven mis- and disinformation campaigns and their impact on democratic discourses. This includes how news media organizations are responding, and which actors/organizations are involved in using AI tools to generate mis- and disinformation, for example, whether this is government actor-driven, amplified by bots or shared by private individuals.

- To study the societal impact of algorithmic design-making, including the operation of content governance tools to understand algorithmic decision-making and auditing processes, and to hold those responsible for deploying them accountable.

- To address the disparity between those who can access and effectively leverage AI systems and those who cannot, that is, the 'AI divide'. The implications of AI systems for democratic participation, especially in the Global Majority World, require further research to avoid deepening this divide. Research is also needed to identify barriers to participation by people from the Global Majority World in developing standards and practices for AI systems.

References

- Abdulrauf, L. A., & Dube, H. (Eds) (2024). *Data Privacy Law in Africa: Emerging Perspectives*. Pretoria University Law Press.
- Adams, R. (2022). *AI in Africa: Key Concerns and Policy Considerations for the Future of the Continent*. Africa Policy Research Institute (APRI), Germany, Policy Brief.
- Ahmad, S., & Greb, M. (2022). Automating social media content moderation: Implications for governance and labour discretion. *Work in the Global Economy*, *2*(2), 176-198. https://doi.org/10.1332/273241721X16647876031174
- Alexander, F. (2024). UN attempts AI power grab. The West is unhappy. CEPA, 24 July. https://cepa.org/article/un-attempts-ai-power-grab-the-west-is-unhappy
- Allen, D., & Weyl, E. G. (2024). The real dangers of generative AI. *Journal of Democracy*, *35*(1), 146-162. www.journalofdemocracy.org/articles/the-real-dangers-of-generative-ai
- Amnesty International, & AI Forensics. (2023). *Driven into the Darkness: How TikTok's 'For You' Feed Encourages Self-Harm and Suicidal Ideation*. www.amnesty.org/en/documents/POL40/7350/2023/en
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973-989. https://doi.org/10.1177/1461444816676645
- Ang, P.-H., & Haristya, S. (2024). The governance, legitimacy and efficacy of Facebook's Oversight Board: A model for global tech platforms? *Emerging Media*, *2*(2). https://doi.org/10.1177/27523543241266860
- Annoni, A., Benczur, P., Bertoldi, P., Delipetrev, B., *et al.* (2018). *Artificial Intelligence: A European Perspective*. European Commission, JRC EUR 29425.
- Avle, S. (2022). Hardware and data in the platform era: Chinese smartphones in Africa. *Media, Culture & Society*, *44*(8), 1473-1489. https://doi.org/10.1177/01634437221128935
- AWS (no date) Just Walk Out technology. https://aws.amazon.com/just-walk-out
- Bachelet, M. (2019). Human rights in the digital age – Can they make a difference? OHCHR Keynote Speech. www.ohchr.org/en/speeches/2019/10/human-rights-digital-age
- Baecker, C., Alabbadi, O., Yogiputra, G. P., & Tien Dung, N. (2023). Threats Provided by Artificial Intelligence that Could Disrupt the Democratic System. Scientific Paper, University of Applied Science, Brandenburg.
- Bail, C. (2021). *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press.
- Bakke, N. A., & Barland, J. (2022). Disruptive innovations and paradigm shifts in journalism as a business: From advertisers first to readers first and traditional operational models to the AI factory. *SAGE Open*, *12*(2), 1-18. https://doi.org/10.1177/21582440221094819
- Bansal, V. (2024). This delivery app takes away health insurance when workers don't meet quotas. Rest of World Reporting Global Tech Stories, 12 April. https://restofworld.org/2024/swiggy-health-insurance-quotas
- Barocas, S., & Nissenbaum, H. (2014). Big Data's End Run Around Anonymity and Consent. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds), *Privacy, Big Data, and the Public Good* (pp. 44-75). Cambridge University Press.
- Bates, J., Kennedy, H., Perea, I. M., Oman, S., & Pinney, L. (2023). Socially meaningful transparency in data-based systems: Reflections and proposals from practice. *Journal of Documentation*, *80*(1), 1-20. https://doi.org/DOI:10.1108/jd-01-2023-0006
- Beckett, C. (2019). *New Powers, New Responsibilities: A Global Survey of Journalism and Artificial Intelligence*. LSE, POLIS (Department of Media and Communications), UK and Google News Initiative.
- Beckett, C., & Yaseen, M. (2023). *Generating Change: A Global Survey of What News Organisations Are Doing with AI*. POLIS (Department of Media and Communications), London School of Economics and Political Science, UK and Google News Initiative.
- Belenguer, L. (2022). AI bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, *2*(4), 771-787. https://doi.org/10.1007/s43681-022-00138-8
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In ACM (Ed.), *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).
- Berger, C., Freihse, C., Kettemann, M. C., Mosene, K., & Hofmann, V. (2023a). The EU Elections 2024: How to build resilience against disinformation campaigns on social platforms. Upgrade Democracy, 28 August. https://upgradedemocracy.de/en/impulse/the-eu-elections-2024-how-to-build-resilience-against-disinformation-campaigns-on-social-platforms
- Berger, C., Freihse, C., Mosene, K., Kettemann, M. C., & Hofmann, V. (2023b). Threats to democracy: Climate misinformation and gendered disinformation. Upgrade Democracy, 26 July. https://upgradedemocracy.de/en/impulse/climate-misinformation-and-gender-related-disinformation-responsibilities-of-civil-society-public-sector-and-the-media
- Berthelot, A., Caron, E., Jay, M., & Lefèvre, L. (2024). Estimating the environmental impact of Generative-AI services using an LCA-based methodology. *31st CIRP Conference on Life Cycle Engineering, 122*, 707-712. https://doi.org/10.1016/j.procir.2024.01.098

- Biju, P. R., & Gayathri, O. (2023). Self-breeding fake news: Bots and artificial intelligence perpetuate social polarization in India's conflict zones. *The International Journal of Information, Diversity, & Inclusion*, *7*(1/2), 1-25. https://doi.org/10.33137/ijidi.v7i1/2.39409
- Birwe, H. (2024). *La désinformation basée sur le genre en contexte de crises: Quels mécanismes de prévention et dispositifs de lutte pour les pays du Sahel?* IDOS Discussion Paper, Géographe diplômé de Sciences Po Lyon et de l'Université Cheikh Anta Diop de Dakar.
- Bitter, A. (2024). Amazon's just walk out: Technology relies on hundreds of workers in India watching you shop. *Business Insider India*, 3 August. www.businessinsider.in/retail/news/amazons-just-walk-out-technology-relies-on-hundreds-of-workers-in-india-watching-you-shop/articleshow/109014034.cms
- Boix, C. (2022). AI and the Economic and Informational Foundations of Democracy. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, *et al.* (Eds), *The Oxford Handbook of AI Governance* (Chapter 35). Oxford University Press.
- Bonfanti, M. E. (2020). *The Weaponisation of Synthetic Media: What Threat Does This Pose to National Security?* Center for Security Studies (CSS) at Zürich, Elcano Royal Institute for International and Strategic Studies (Real Instituto Elcano).
- Bontcheva, K., Papdopoulous, S., Tsalakanidou, F., Dutkiewicz, L., *et al.* (2024). *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*. Vera.ai, AI4Trust, AI4Media, & TITAN. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf
- Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, *3*(e32), 1-21. https://doi.org/10.1017/dap.2021.20
- Borocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.
- Bradshaw, S., & Howard, P. N. (2019). *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*. Oxford Internet Institute, University of Oxford Project on Computational Propaganda.
- Bradshaw, S., Bailey, H., & Howard, P. N. (2021). *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*. Oxford Internet Institute, University of Oxford Computational Propaganda Research Project.
- Bradshaw, S., Campbell-Smith, U., Henle, A., Perini, A., *et al.* (2020). *Country Case Studies, Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*. Oxford Internet Institute, University of Oxford Project on Computational Propaganda. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/03/Case-Studies_FINAL.pdf
- Breuer, A. (2024). *Information Integrity and Information Pollution: Vulnerabilities and Impact on Social Cohesion and Democracy in Mexico*. IDOS Discussion Paper 2/2024. IDOS (German Institute of Development and Sustainability). www.idos-research.de/uploads/media/DP_2.2024.pdf
- Brevini, B. (2021). *Is AI Good for the Planet?*. Polity Press.
- Brewster, T. (2024). Musk's X fired 80% of engineers working on trust and safety, Australian Government says. *Forbes*, 10 January. www.forbes.com/sites/ssbrewster/2024/01/10/elon-musk-fired-80-per-cent-of-twitter-x-engineers-working-on-trust-and-safety
- Broniatowski, D. A., Simons, J. R., Gu, J., Jamison, A. M., & Abroms, L. C. (2023). The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. *Science Advances*, *9*(37), 1-17. https://doi.org/10.1126/sciadv.adh2132
- Bullock, J. B., Chen, Y.-C., Himmelreich, J., Hudson, V. M., *et al.* (2022). *The Oxford Handbook of AI Governance*. Oxford University Press.
- Burgess, M. (2016). Google's DeepMind trains AI to cut its energy bills by 40%. Wired, 28 July. www.wired.com/story/google-deepmind-data-centres-efficiency
- Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245-317. https://doi.org/10.1613/jair.1.12228
- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1-12. https://journals.sagepub.com/doi/epub/10.1177/2053951715622512
- Calice, M. N., Bao, L., Freiling, I., Howell, E., *et al.* (2023). Polarized platforms? How partisanship shapes perceptions of 'algorithmic news bias'. *New Media & Society*, *25*(11), 2833-2854. https://doi.org/10.1177/14614448211034159
- Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. ArXiv. https://doi.org/10.48550/arXiv.2010.04053
- Chi, N., Lurie, E., & Mulligan, D. K. (2021). Reconfiguring diversity and inclusion for AI ethics. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society* (pp. 447-457). https://doi.org/10.1145/3461702.3462622
- Chouliaraki, L., & Georgiou, M. (2022). *The Digital Border: Migration, Technology, Power*. NYU Press.
- Christodoulou, E., & Iordanou, K. (2021). Democracy under attack: Challenges of addressing ethical issues of AI and big data for more democratic digital media and societies. *Frontiers in Political Science*, *3*, 1-17. www.frontiersin.org/articles/10.3389/fpos.2021.682945/full
- Christopher, N. (2021). Amazon's 'Delivery Premier League' gamifies gig work in India. Rest of World Reporting Global Tech Stories, 25 October. https://restofworld.org/2021/amazons-delivery-premier-league-gamifies-gig-work-in-india
- Cooke, D. (2023). *Synthetic Media and Election Integrity: Defending Our Democracies*. Alan Turing Institute, Centre for Emerging Technology and Security (CETaS).
- Coombs, W. T., & Holladay, S. J. (Eds) (2022). *The Handbook of Crisis Communication*, Second Edition. John Wiley & Sons.

- Council of Europe. (2018). *Recommendation CM/Rec(2018)1 of the Committee of Ministers to Member States on Media Pluralism and Transparency of Media Ownership*. CM/Rec(2018)1.
- Council of Europe. (2023). *Guidelines for the Responsible Implementation of Artificial Intelligence (AI) in Journalism*, https://www.coe.int/en/web/freedom-expression/-/guidelines-on-the-responsible-implementation-of-artificial-intelligence-ai-systems-in-journalism
- Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2023). The AI gambit: Leveraging artificial intelligence to combat climate change – Opportunities, challenges, and recommendations. *AI & SOCIETY*, *38*(1), 283-307. https://doi.org/10.1007/s00146-021-01294-x
- Crosset, V., & Dupont, B. (2022). Cognitive assemblages: The entangled nature of algorithmic content moderation. *Big Data & Society*, *9*(2), 1-13. https://doi.org/10.1177/20539517221143361
- De Gregorio, G. (2023). The normative power of artificial intelligence. *Indiana Journal of Global Legal Studies, 30*(2), 55-80. https://ssrn.com/abstract=4436287
- De Gregorio, G., & Dunn, P. (2023). Artificial Intelligence and Freedom of Expression. In A. Quintavalla & J. Temperman (Eds), *Artificial Intelligence and Human Rights* (pp. 76-90). Oxford University Press.
- De Gregorio, G., & Stremlau, N. (2023). Inequalities and content moderation. *Global Policy*, *14*(5), 870-879. https://doi.org/10.1111/1758-5899.13243
- de Vivo, I. (2023). The 'neo-intermediation' of large on-line platforms: Perspectives of analysis of the 'state of health' of the digital information ecosystem. *Communications*, *48*(3), 420-439. https://doi.org/10.1515/commun-2022-0102
- de Lima Santos, M.-F., Yeung, W. N., & Dodds, T. (2024). Guiding the way: A comprehensive examination of AI guidelines in global media. *AI & SOCIETY*, 1-19. https://doi.org/10.1007/s00146-024-01973-5
- Dias Oliva, T. (2020). Content moderation technologies: Applying human rights standards to protect freedom of Eepression. *Human Rights Law Review*, *20*(4), 607-640. https://doi.org/10.1093/hrlr/ngaa032
- Dierickx, L., Lindén, C.-G., & Opdahl, A. L. (2023a). Automated fact-checking to support professional practices: Systematic literature review and meta-analysis. *International Journal of Communication*, *17*(2023), 5170-5190. https://ijoc.org/index.php/ijoc/article/view/21071
- Dierickx, L., Linden, C.-G., Opdahl, A., Khan, S. A., & Rojas, D. (2023b). AI in the newsroom: A data quality assessment framework for employing machine learning in journalistic workflows. *Proceedings of the 5th International Conference on Advanced Research Methods and Analytics*. http://hdl.handle.net/10251/201692
- Douek, E. (2024). The Meta Oversight Board and the empty promise of legitimacy. *Harvard Journal of Law & Technology, 37*(2), 373-445. http://dx.doi.org/10.2139/ssrn.4565180
- Dutta, S., & Lanvin, B. (2023). *Network Readiness Index 2023: Trust in a Network Society: A Crisis of the Digital Age?* Portulans Institute, University of Oxford.
- EC (European Commission). (2016b). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*. OJ L 119/1. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
- EC. (2022b). *European Commission Digital Strategy*. European Commission C(2022) 4388 final. https://commission.europa.eu/publications/european-commission-digital-strategy_en
- EC. (2024a). Commission requests information from X on decreasing content moderation resources under the Digital Services Act. European Commission News, 8 May. https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-x-decreasing-content-moderation-resources-under-digital-services
- EC. (2024c). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689
- EC. (2024d). Funded projects in the fight against disinformation. https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation/funded-projects-fight-against-disinformation_en
- Eichler, H. (2023). Cross-Media and Platformised Journalism: ARD's Innovation Attempts at Becoming a Multi-Platform Contributor to Society. In M. Puppis & C. Ali (Eds), *Public Service Media's Contribution to Society: RIPE@2021* (pp. 265-288). Nordicom.
- Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society*, *7*(2), 1-13. https://doi.org/10.1177/2053951720932296
- Elliott, V. (2024). AI-fakes detection is failing voters in the Global South. Wired, 2 September. www.wired.com/story/generative-ai-detection-gap
- eSafety Commissioner. (2024a). Report reveals the extent of deep cuts to safety staff and gaps in Twitter/X's measures to tackle online hate. Media Release, 11 January. www.esafety.gov.au/newsroom/media-releases/report-reveals-the-extent-of-deep-cuts-to-safety-staff-and-gaps-in-twitter/xs-measures-to-tackle-online-hate
- eSafety Commissioner. (2024b). *Basic Online Safety Expectations: Summary of response from X Corp. (Twitter) to eSafety's Transparency Notice on Online Hate*. eSafetyCommissioner, Australia.
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St Martin's Publishing Group.
- Evans, R., Jackson, D., & Murphy, J. (2023). Google News and machine gatekeepers: Algorithmic personalisation and news diversity in online news search. *Digital Journalism*, *11*(9), 1682-1700. https://doi.org/10.1080/21670811.2022.2055596

- Fatafta, M. (2024). How Meta censors Palestinian voices. Access Now, 19 February. www.accessnow.org/publication/how-meta-censors-palestinian-voices
- Fendji, J. L. K. E. (2024). From left behind to left out: Generative AI or the next pain of the unconnected. *Harvard Data Science Review*, 1-3. https://hdsr.mitpress.mit.edu/pub/jo4xfs0s/release/2
- Ferrara, E. (2024a). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 1-15. https://doi.org/10.3390/sci6010003
- Fertmann, M., & Kettemann, M. C. (Eds) (2021). *Viral Information: How States and Platforms Deal with Covid-19-Related Disinformation, An Exploratory Study of 20 Countries*. Verlag Hans-Bredow-Institut.
- Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., *et al*. (2023). Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, *7*(11), 1818-1821. https://doi.org/10.1038/s41562-023-01726-2
- Fischer-Lescano, A. (2016). Struggles for a global Internet constitution: Protecting global communication structures against surveillance measures. *Global Constitutionalism*, *5*(2), 145-172. https://doi.org/10.1017/S204538171600006X
- Floridi, L., & Nobre, A. C. (2024). Anthropomorphising machines and computerising minds: The crosswiring of languages between artificial intelligence and brain & cognitive sciences. *Minds and Machines*, *34*(5), 1-9. https://doi.org/10.2139/ssrn.4738331
- Fontes, C., Hohma, E., Corrigan, C. C., & Lütge, C. (2022). AI-powered public surveillance systems: Why we (might) need them and how we want them. *Technology in Society*, *71*(2022), 1-12. https://doi.org/10.1016/j.techsoc.2022.102137
- Forum on Information and Democracy. (2024a). *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*. https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf
- Forum on Information and Democracy. (2024b). Data protection authorities stand up to big tech: The importance of democratic institutions to safeguard citizens' privacy rights. Unpacking Current Developments in the Information Space, Insight Nr. 1. https://informationdemocracy.org/wp-content/uploads/2024/09/FID-Insight-Nr-1-Data-protection-authorities-and-AI.pdf
- Frau-Meigs, D. (2024b). On AI Pedagogy. Presentation to VOICES – European Festival of Journalism and Media Literacy, March.
- Funk, A., Shahbaz, A., & Vesteinsson, K. (2023). Freedom on the Net 2023: *The Repressive Power of Artificial Intelligence*. Freedom House. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence
- Galli, F., Loreggia, A., & Sartor, G. (2023). The Regulation of Content Moderation. In D. Moura Vicente, S. de Vasconcelos Casimiro, & C. Chen (Eds), *The Legal Challenges of the Fourth Industrial Revolution* (pp. 63-87). Springer International Publishing.
- Geiger, C. (2024). Elaborating a human rights-friendly copyright framework for generative AI. *IIC – International Review of Intellectual Property and Competition Law*, *2024*(3 June), 1-37. https://doi.org/10.1007/s40319-024-01481-5
- Geiß, S., Magin, M., Jürgens, P., & Stark, B. (2021). Loopholes in the echo chambers: How the echo chamber metaphor oversimplifies the effects of information gateways on opinion expression. *Digital Journalism*, *9*(5), 660-686. https://doi.org/10.1080/21670811.2021.1873811
- Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, A. S. (2023). Towards possibilities & impossibilities of AI-generated text detection: A survey. *ArXiv*. https://doi.org/10.48550/arXiv.2310.15264
- Gil de Zúñiga, H., Borah, P., & Goyanes, M. (2021). How do people learn about politics when inadvertently exposed to news? Incidental news paradoxical direct and indirect effects on political knowledge. *Computers in Human Behavior*, *121*, 1-9. www.sciencedirect.com/science/article/pii/S0747563221001266?casa_token=e0jAerFl9RkAAAAA:KkK-Ad1MxaQB9UVbkQcaAsRNKnofBOZP_C0A0YXkqV8j6satCJLfeKJoThF7xGxmvMxVH2Tf
- Gil de Zúñiga, H., Goyanes, M., & Durotoye, T. (2023). A scholarly definition of artificial intelligence (AI): Advancing AI as a conceptual framework in communication research. *Political Communication*, *41*(2), 317-334. https://doi.org/10.1080/10584609.2023.2290497
- Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 1-5. https://doi.org/10.1177/2053951720943234
- Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., *et al*. (2023). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review, 9*(4).https://ssrn.com/abstract=4459448
- Goldman Sachs. (2024). AI is poised to drive 160% increase in data center power demand. Goldman Sachs. 14 May. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand
- Google. (2022). Our commitment to climate-conscious data center cooling. 21 November. https://blog.google/outreach-initiatives/sustainability/our-commitment-to-climate-conscious-data-center-cooling
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1-15. https://doi.org/10.1177/2053951719897945
- Gulati, R. (2023). Meta's Oversight Board and transnational hybrid adjudication – What consequences for international law? *German Law Journal*, *24*(3), 473-493. https://doi.org/10.1017/glj.2023.34
- Gunkel, D. J. (Ed.). (2024). *Handbook on the Ethics of Artificial Intelligence*. Edward Elgar Publishing.
- Gurumurthy, A., & Chami, N. (2019). *The Wicked Problem of AI Governance*. IT for Change, India and Friedrich Ebert Stiftung.
- Haas, J., & Kettemann, M. (2024). *Platform and Content Governance in Times of Crisis*. OSCE (Organization for Security and Co-operation in Europe).

- Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review*, *47*(7), 1264-1279. https://doi.org/10.1108/OIR-08-2021-0452
- Harris, J. (2023). 'There was all sorts of toxic behaviour': Timnit Gebru on her sacking by Google, AI's dangers and big tech's biases. *The Guardian*, 22 May. www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases
- Hase, V., Boczek, K., & Scharkow, M. (2023). Adapting to affordances and audiences? A cross-platform, multi-modal analysis of the platformization of news on Facebook, Instagram, TikTok, and Twitter. *Digital Journalism*, *11*(8), 1499-1520. https://doi.org/10.1080/21670811.2022.2128389
- Hasimi, L., & Poniszewska-Marańda, A. (2024). Detection of disinformation and content filtering using machine learning: Implications to human rights and freedom of speech. ROMCIR 2024: The 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2024, the 46th European Conference on Information Retrieval). https://ceur-ws.org/Vol-3677/paper6.pdf
- Heeks, R. (2022). Digital inequality beyond the digital divide: Conceptualizing adverse digital incorporation in the global South. *Information Technology for Development*, *28*(4), 688-704. https://doi.org/10.1080/02681102.2022.2068492
- Heitz, L., Rozgonyi, K., & Kostic, B. (2021). AI in Content Curation and Media Pluralism. In D. Wagner & J. Haas (Eds), *Spotlight on Artificial Intelligence and Freedom of Expression – A Policy Manual* (pp. 56-70). OSCE (Organization for Security and Co-operation in Europe).
- Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207. https://doi.org/https://doi.org/10.1080/1369118X.2016.1271900
- Helberger, N., Van Drunen, M., Eskens, S., & Bastian, M. (2020). A freedom of expression perspective on AI in the media – with a special focus on editorial decision making on social media platforms and in the news media. *European Journal of Law and Technology*, *11*(3), 1-28. https://ejlt.org/index.php/ejlt/article/view/752
- Helberger, N., Van Drunen, M., Moeller, J., Vrijenhoek, S., & Eskens, S. (2022). Towards a normative perspective on journalistic AI: Embracing the messy reality of normative ideals. *Digital Journalism*, *10*(10), 1605-1626. https://doi.org/10.1080/21670811.2022.2152195
- Helming, C. (2023). Microsoft's Bing Chat: A source of misinformation on elections. *Algorithm Watch*, 15 December. https://algorithmwatch.org/en/microsofts-bing-source-misinformation-elections
- Horwitz, J. (2021). The Facebook files. *Wall Street Journal*. www.wsj.com/articles/the-facebook-files-11631713039
- Horowitz, M., Milosavljević, M., & Van den Bulck, H. (2022). The Use of Artificial Intelligence by Public Service Media: Between Advantages and Threats. In C. El Morr (Ed.), *AI and Society: Tensions and Opportunities* (pp. 14). Chapman and Hall/CRC.
- HRW (Human Rights Watch). (2023). *Meta's Broken Promises: Systemic Censorship of Palestine Content on Instgram and Facebook*. www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and
- Iamnitchi, A., Hall, L. O., Horawalavithana, S., Mubang, F., Ng, K. W., & Skvoretz, J. (2023). Modeling information diffusion in social media: Data-driven observations. *Frontiers in Big Data*, *6*, 1-19. https://doi.org/10.3389/fdata.2023.1135191
- iea50. (2024). Data centres and data transmission networks. IEA (International Energy Agency). www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks
- IEEE (Institute of Electrical and Electronics Engineers). (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*. Version 2. https://ieeexplore.ieee.org/document/9398613
- Innerarity, D. (2024). *Artificial Intelligence and Democracy*. UNESCO Regional Office, Montevideo and CLACSO (Latin American Council of Social Sciences). https://cadmus.eui.eu/bitstream/handle/1814/77333/AI_democracy_2024.pdf?sequence=1
- International IDEA (Institute for Democracy and Electoral Assistance). (2023). *The Global State of Democracy 2023: The New Checks and Balances*. www.idea.int/gsod/2023
- Ivanova, I. (2022). These formerly banned Twitter accounts have been reinstated since Elon Musk took over. CBS News, 21 November. www.cbsnews.com/news/twitter-accounts-reinstated-elon-musk-donald-trump-kanye-ye-jordan-peterson-kathy-griffin-andrew-tate
- Jagannatha, A., Rawat, B. P. S., & Yu, H. (2021). Membership inference attack susceptibility of clinical language models. *ArXiv*. https://doi.org/10.48550/arXiv.2104.08305
- Jin, Y., & Austin, L. (Eds) (2022). *Social Media and Crisis Communication*. Routledge.
- Johnson, G. M. (2023). Are algorithms value-free? Feminist theoretical virtues in machine learning. *Journal of Moral Philosophy*, *21*(1-2), 27-61. https://doi.org/10.1163/17455243-20234372
- Jørgensen, R. F. (2017). What platforms mean when they talk about human rights. *Policy & Internet*, *9*(3), 280-296. https://doi.org/10.1002/poi3.152
- Jungherr, A., & Schroeder, R. (2023). Artificial intelligence and the public arena. *Communication Theory*, *33*(2-3), 164-173. https://academic.oup.com/ct/article/33/2-3/164/7202294
- Katzenbach, C. (2021). 'AI will fix this' – The technical, discursive, and political turn to AI in governing communication. *Big Data & Society*, *8*(2), 1-8. https://doi.org/10.1177/20539517211046182
- Keller, T. R., & Klinger, U. (2019). Social bots in election campaigns: Theoretical, empirical, and methodological implications. *Political Communication*, *36*(1), 171-189. https://doi.org/10.1080/10584609.2018.1526238

- Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, *29*(1-4), 55-81. https://doi.org/10.1163/18750230-02901005
- Kettemann, M. C., & Lachmayer, K. (Eds) (2021). *Pandemocracy in Europe: Power, Parliaments and People in Times of COVID-19*. Bloomsbury.
- Kettemann, M. C., & Schulz, W. (2023). *Platform://Democracy – Perspectives on Platform Power, Public Values and the Potential of Social Media Councils*. Verlag Hans-Bredow-Institut.
- Kim, K., & Moon, S.-I. (2021). When algorithmic transparency failed: Controversies over algorithm-driven content curation in the South Korean digital environment. *American Behavioral Scientist*, *65*(6), 847-862. https://doi.org/10.1177/0002764221989783
- Kostygina, G., Kim, Y., Seeskin, Z., LeClere, F., & Emery, S. (2023). Disclosure standards for social media and generative artificial intelligence research: Toward transparency and replicability. *Social Media + Society*, *9*(4), 1-12. https://doi.org/10.1177/20563051231216947
- Kothari, A., & Cruikshank, S. A. (2022). Artificial intelligence and journalism: An agenda for journalism research in Africa. *African Journalism Studies*, *43*(1), 17-33. https://doi.org/10.1080/23743670.2021.1999840
- Kreiss, D., & McGregor, S. C. (2023). A review and provocation: On polarization and platforms. *New Media & Society*, *26*(1), 556-579. https://doi.org/10.1177/14614448231161880
- Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104-117. https://doi.org/10.1017/XPS.2020.37
- Kulesz, M. (2018). *Culture, Platforms and Machines: The Impact of Artificial Intelligence on the Diversity of Cultural Expressions*. UNESCO Report to Intergovernmental Committee for the Protection and Promotion of the Diversity of Cultural Expressions, DCE/18/12.IGC/INF.4.
- Lee, J. (2024). CCPA/CPRA: Consumers bear the burden as companies bear the crown. *Hastings International & Comparative Law Review,*, *47*(2), 129-152. https://repository.uclawsf.edu/hastings_international_comparative_law_review/vol47/iss2/5
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611-627. https://doi.org/10.1007/s13347-017-0279-x
- Lin, B., & Lewis, S. C. (2022). The one thing journalistic AI just might do for democracy. *Digital Journalism*, *10*(10), 1627-1649. https://doi.org/10.1080/21670811.2022.2084131
- Liu, Y., Zhang, K., Li, Y., Yan, Z., *et al*. (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models. ArXiv. https://doi.org/10.48550/arXiv.2402.17177
- Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., *et al*. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *ArXiv*. https://doi.org/10.1016/j.inffus.2024.102301
- Macdonald, S., Correia, S. G., & Watkin, A.-L. (2019). Regulating terrorist content on social media: Automation and the rule of law. *International Journal of Law in Context*, *15*(2), 183-197. https://doi.org/10.1017/S1744552319000119
- Madrid-Morales, D., & Wasserman, H. (2022). Research methods in comparative disinformation studies. In H. Wasserman & D. Madrid-Morales (Eds.), *Disinformation in the Global South* (pp. 41-57). Wiley Blackwell.
- Mahler, T. (2022). Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal. In L. Colonna & S. Greenstein (Eds), *Law in the Era of Artificial Intelligence: Nordic Yearbook of Law and Informatics 2020-2021* (pp. 247-270). The Swedish Law and Informatics Research Institute.
- Makan, N. (2023). Sustainable industrial machine learning and artificial intelligence: A framework for environmentally responsible innovation. *Transactions on Recent Developments in Artificial Intelligence and Machine Learning*, *15*(15). https://journals.threws.com/index.php/TRDAIML/article/view/184
- Makhortykh, M., Urman, A., Münch, F. V., Heldt, A., Dreyer, S., & Kettemann, M. C. (2022). Not all who are bots are evil: A cross-platform analysis of automated agent governance. *New Media & Society*, *24*(4), 964-981. https://doi.org/10.1177/14614448221079035
- Marconi, F. (2020). *Newsmakers – Artificial Intelligence and the Future of Journalism*. Columbia University Press.
- Marinova, Z., Spangenberg, J., Teyssou, D., Papadopoulos, S., et al. (2020). Weverify: Wider and Enhanced Verification for You Project Overview and Tools. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. https://ieeexplore.ieee.org/document/9106056
- Martin, K. (2022). Google Research: Who Is Responsible for Ethics of AI? In K. Martin (Ed.), *Ethics of Data and Analytics: Concepts and Cases* (pp. 13). Auerbach Publications.
- Masur, P. K. (2020). How online privacy literacy supports self-data protection and self-determination in the age of information. *Media and Communication*, *8*(2), 1-12. www.ssoar.info/ssoar/handle/document/69359
- Meese, J., & Hurcombe, E. (2021). Facebook, news media and platform dependency: The institutional impacts of news distribution on social platforms. *New Media & Society*, *23*(8), 2367-2384. https://doi.org/10.1177/1461444820926472
- Michael, A. (2023). Artificial intelligence, democracy and elections. European Parliament Briefing. www.europarl.europa.eu/RegData/etudes/BRIE/2023/751478/EPRS_BRI(2023)751478_EN.pdf
- Microsoft. (2023). *Diversity and Inclusion Report 2023: A Decade of Transparency, Commitment and Progress*.

- Miguel, R., & Krack, N. (2023). Platforms' policies on AI-manipulated and generated misinformation. EU DisinfoLab, 28 September. www.disinfo.eu/publications/platforms-policies-on-ai-manipulated-and-generated-misinformation
- Milmo, D. (2024). OpenAI putting 'shiny products' above safety, says departing researcher. *The Observer*, 28 May. www.theguardian.com/technology/article/2024/may/18/openai-putting-shiny-products-above-safety-says-departing-researcher
- Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. *AI and Ethics*, *2023*, 1-31. https://doi.org/10.1007/s43681-023-00289-2
- Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959-977. https://doi.org/10.1080/1369118X.2018.1444076
- Monasterio Astobiza, A., Ausín, T., Liedo, B., Toboso, M., Aparicio, M., & López, D. (2022). Ethical Governance of AI in the Global South: A Human Rights Approach to Responsible Use of AI. *Proceedings 81*(136), 1-5. https://doi.org/10.3390/proceedings2022081136
- Moran, R. E., & Shaikh, S. J. (2022). Robots in the news and newsrooms: Unpacking Meta-journalistic discourse on the use of artificial intelligence in journalism. *Digital Journalism*, *10*(10), 1756-1774. https://doi.org/10.1080/21670811.2022.2085129
- Motta, M., Hwang, J., & Stecula, D. (2024). What goes down must come up? Pandemic-related misinformation search behavior during an unplanned Facebook outage. *Health Communication*, *39*(10), 2041-2052. https://doi.org/10.1080/10410236.2023.2254583
- Müller, M., & Kettemann, M. C. (2024). European Approaches to the Regulation of Digital Technologies. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, *et al.* (Eds), *Introduction to Digital Humanism: A Textbook* (pp. 623-637). Springer Nature Switzerland.
- Muñoz, K. (2023). *The Transformative Role of AI in Reshaping Electoral Politics*. DGAP Memo, No. 4. Deutsche Gesellschaft für Auswärtige Politik e.V.
- Mutsvairo, B., & Ragnedda, M. (Eds) (2019). *Mapping Digital Divide in Africa: A Mediated Analysis*. Amsterdam University Press.
- Naeem, S. B., Bhatti, R., & Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, *38*(2), 143-149. https://doi.org/10.1111/hir.12320
- Nagaraj, A. (2020). Under watch: Indian city workers protest digital surveillance. Reuters, 17 March. www.reuters.com/article/idUSKBN21404S
- Nahmias, Y., & Perel, M. (2021). The oversight of content moderation by AI: Impact assessments and their limitations. *Harvard Journal on Legislation*, *58*(1), 145-194. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3565025
- Ning, L., Liu, L., Wu, J., Wu, N., Berlowitz, D., Prakash, S., Green, B., O'Banion, S., & Xie, J. (2024). *User-LLM: Efficient LLM contextualization with user embeddings*. arXiv. https://doi.org/10.48550/arXiv.2402.13598
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- Nordic Council of Ministers Secretariat. (2023). *A Nordic Approach to Democratic Debate in the Age of Big Tech : Recommendations from the Nordic Think Tank for Tech and Democracy*. http://dx.doi.org/10.6027/nord2023-004
- Nowotny, H. (2021). *In AI We Trust: Power, Illusion and Control of Predictive Algorithms*. Polity Press.
- O'Brien, I. (2024). Data center emissions probably 662% higher than big tech claims. Can it keep up the ruse? *The Guardian*, 15 September. www.theguardian.com/technology/2024/sep/15/data-center-gas-emissions-tech
- O'Brien, M., & Swenson, A. (2024). Tech companies sign accord to combat AI-generated election trickery. The Associated Press, 16 February. https://apnews.com/article/ai-generated-election-deepfakes-munich-accord-meta-google-microsoft-tiktok-x-c40924ffc68c94fac74fa994c520fc06
- OECD (Organisation for Economic Co-operation and Development). (2022a). *Recommendation of the Council on Artificial Intelligence*. https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf
- OECD. (2023). Catalogue of tools & metrics for trustworthy AI. https://oecd.ai/en/catalogue
- Ofcom. (2023d). *Online Nation 2023 Report*. www.ofcom.org.uk/media-use-and-attitudes/online-habits/online-nation
- Offenhuber, D. (2024). Shapes and frictions of synthetic data. *Big Data & Society*, *11*(2), 1-16. https://doi.org/10.1177/20539517241249390
- OHCHR (Office of the United Nations High Commissioner for Human Rights). (1993). *Vienna Declaration and Programme of Action*. www.ohchr.org/en/instruments-mechanisms/instruments/vienna-declaration-and-programme-action
- Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., *et al.* (2024). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, *18*(2), 124-141. www.tandfonline.com/doi/full/10.1080/19312458.2023.2181319
- Okolo, C. T. (2023). AI in the Global South: Opportunities and challenges towards more inclusive governance. Brookings Institution Commentary. 1 November. www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance
- Ong, J. C. L., Chang, S. Y.-H., William, W., Butte, A. J., *et al.* (2024). Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, *6*(6), e428-e432. https://doi.org/10.1016/S2589-7500(24)00061-X
- Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, *9*(2), 1-13. https://doi.org/10.1177/20539517221115666

- Park, Y. J. (2024). Algorithmic bias, marketplaces, and diversity regulation. In Proceedings of the TPRC 2024 Conference, Washington DC. https://ssrn.com/abstract=4912448
- Pasquinelli, M. (2023). *The Eye of the Master: A Social History of Artificial Intelligence*. Verso Books.
- Paul, R., Carmel, E., & Cobbe, J. (Eds) (2024). *Handbook on Public Policy and Artificial Intelligence*. Edward Elgar Publishing.
- Pfeiffer, J., Gutschow, J., Haas, C., Möslein, F., *et al*. (2023). Algorithmic fairness in AI. *Business & Information Systems Engineering*, *65*(2), 209-222. https://doi.org/10.1007/s12599-023-00787-x
- Pocino, P. V. (2021). *Algorithms in the Newsrooms: Challenges and Recommendations for Artificial Intelligence with the Ethical Values of Journalism*. Catalan Press Council.
- Poell, T., Nieborg, D. B., & Duffy, B. E. (2023). Spaces of negotiation: Analyzing platform power in the news industry. *Digital Journalism*, *11*(8), 1391-1409. https://doi.org/10.1080/21670811.2022.2103011
- Pollicino, O., & De Gregorio, G. (2022). Constitutional democracy, platform powers and digital populism. *Constitutional Studies*, *8*(1), 11-34. https://constitutionalstudies.wisc.edu/index.php/cs/article/view/87
- Posetti, J., & Bontcheva, K. (2020). *Disinfodemic: Deciphering COVID-19 Disinformation*. UNESCO Policy Brief 1. https://unesdoc.unesco.org/ark:/48223/pf0000374416
- PublicSpaces International. (2024). How we think about AI is largely dictated by Big Tech. https://english.publicspaces.net/2024/06/03/how-we-think-about-ai-is-largely-dictated-by-big-tech
- Puddephatt, A. (2021). *Letting the Sun Shine In: Transparency and Accountability in the Digital Age*. UNESCO.
- Pyo, J. Y. (2022). Different stakes, different struggles, and different practices to survive: News organizations and the spectrum of platform dependency. *New Media & Society*, *26*(8), 4572-4588. https://doi.org/10.1177/14614448221123279
- Quintavalla, A., & Temperman, J. (Eds) (2023). *Artificial Intelligence and Human Rights*. Oxford University Press.
- Rajkumar, M., & Ashraf, M. M. (2023). *Response to Draft Amendment to the IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021*. IT for Change, India.
- Ranaivoson, H., Afilipoaie, A., & Domazetovikj, N. (2022). *Media Pluralism in the EU: A Prospective Look at the European Media Freedom Act*. Policy Brief No. 64. Vrije Universiteit Brussel, Studies in Media Innovation Technology Research Group.
- Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research*, *291*(3), 906-917. https://doi.org/10.1016/j.ejor.2020.09.020
- Repucci, S. & Slipowitz, A. (2022). *The Global Expansion of Authoritarian Rule*. Freedom House. https://freedomhouse.org/report/freedom-world/2022/global-expansion-authoritarian-rule
- Ricaurte, P. (2022). Ethics for the majority world: AI and the question of violence at scale. *Media, Culture & Society*, *44*(4), 726-745. https://doi.org/10.1177/01634437221099612
- Richards, N., & Hartzog, W. (2019). The pathologies of digital consent. *Washington University Law Review*, *96*, 1461-1503. https://openscholarship.wustl.edu/cgi/viewcontent.cgi?article=6460&context=law_lawreview
- Rieder, B., & Sire, G. (2014). Conflicts of interest and incentives to bias: A microeconomic critique of Google's tangled position on the Web. *New Media & Society*, *16*(2), 195-211. https://doi.org/10.1177/1461444813481195
- Roche, C., Wall, P. J., & Lewis, D. (2023). Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI and Ethics*, *3*(4), 1095-1115. https://doi.org/10.1007/s43681-022-00218-9
- Ross Arguedas, A., & Simon, F. M. (2023). *Automating Democracy: Generative AI, Journalism, and the Future of Democracy*. Balliol College, Oxford Internet Institute and Institute for Ethics in AI.
- Ross Arguedas, A., Robertson, C., Fletcher, R., & Nielsen, R. (2022a). *Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review*. Reuters Institute for the Study of Journalism, University of Oxford and The Royal Society.
- Ruffo, G., Semeraro, A., Giachanou, A., & Rosso, P. (2023). Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer Science Review*, *47*(2023), 1-26. https://doi.org/10.1016/j.cosrev.2022.100531
- Ruggie, J. G. (2011). *Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises: Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*. United Nations General Assembly A/HRC/17/31. https://digitallibrary.un.org/record/705860?v=pdf
- Ryan-Mosley, T. (2023). How generative AI is boosting the spread of disinformation and propaganda. MIT Technology Review, 4 October. www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/
- Samoilenko, S. A., & Suvorova, I. (2023). Artificial Intelligence and Deepfakes in Strategic Deception Campaigns: The US and Russian Experiences. In E. Pashentsev (Ed.), *The Palgrave Handbook of Malicious Use of AI and Psychological Security* (pp. 507-529). Springer International Publishing.
- Samoili, S., Lopez, C. M., Gomez, G. E., De, P. G., Martinez-Plumed, F., & Delipetrev, B. (2020). *AI WATCH. Defining Artificial Intelligence*. European Commission, JRC EUR30117.
- Sančanin, B., & Penjišević, A. (2022). Use of artificial intelligence for the generation of media content. *Social Informatics Journal*, *1*(1), 1-7. https://doi.org/10.58898/sij.v1i1.01-07
- Schaake, M., & Fukuyama, F. (Eds) (2023). *Digital Technologies in Emerging Countries*. Stanford Cyber Policy Center.
- Schaetz, N., Gagrčin, E., Toth, R., & Emmer, M. (2023). Algorithm dependency in platformized news use. *New Media & Society*, *Online First*, 1-18. https://doi.org/10.1177/14614448231193093

- Schippers, B. (2020). Artificial intelligence and democratic politics. *Political Insight*, *11*(1), 32-35. https://journals.sagepub.com/doi/full/10.1177/2041905820911746
- Schirch, L. (2021). Digital Information, Conflict and Democracy. In L. Schirch (Ed.), *Social Media Impacts on Conflict and Democracy*. Routledge.
- Settle, J. E. (2018). *Frenemies: How Social Media Polarizes America*. Cambridge University Press.
- Simon, F. M. (2022). Uneasy bedfellows: AI in the news, platform companies and the issue of journalistic autonomy. *Digital Journalism*, *10*(10), 1832-1854. https://doi.org/10.1080/21670811.2022.2063150
- Simon, F. M. (2024). *Artificial Intelligence in the News: How AI Retools, Rationalizes, and Reshapes Journalism and the Public Arena*. Tow Center for Jouralism, Oxford Internet Institute, University of Oxford.
- Simon, F. M., & Isaza-Ibarra, L. F. (2023). *AI in the News: Reshaping the Information Ecosystem?* Balliol College, Torch, Reuters Institute, University of Oxford.
- Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School (HKS) Misinformation Review*, *4*(5), 1-11. https://doi.org/10.37016/mr-2020-127
- Sîrbu, A., Pedreschi, D., Giannotti, F., & Kertész, J. (2019). Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PLOS ONE*, *14*(3), 1-20. https://doi.org/10.1371/journal.pone.0213246
- Smith, J. R. (2019). IBM Research releases 'Diversity in Faces' dataset to advance study of fairness in facial recognition systems. Phys.org, 29 January. https://phys.org/news/2019-01-ibm-diversity-dataset-advance-fairness.html
- Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, *9*(26), 1-9. https://doi.org/10.1126/sciadv.adh1850
- Splichal, S. (2022a). *Datafication of Public Opinion and the Public Sphere: How Extraction Replaced Expression of Opinion*. Anthem Press.
- Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2), 1-5. https://doi.org/10.1177/20539517231206794
- Teyssou, D., Leung, J.-M., Apostolidis, E., Apostolidis, K., *et al*. (2017). The InVID Plug-in: Web Video Verification on the Browser. In ACM (Association for Computing Machinery) (Ed.), *Proceedings of the First International Workshop on Multimedia Verification* (pp. 23-30). https://doi.org/10.1145/3132384.3132387
- Thomas, E. (2022). *Conspiracy Clickbait: This One Weird Trick Will Undermine Democracy*. ISD Institute for Strategic Studies.
- Turow, J., Lelkes, Y., Draper, N., & Waldman, A. E. (2023). *Americans Can't Consent to Companies' Use of Their Data: They Admit They Don't Understand It, Say They're Helpless to Control It, and Believe They're Harmed When Firms Use Their Data – Making What Companies Do Illegitimate*. Annenberg School for Communications, University of Pennsylvania. www.asc.upenn.edu/sites/default/files/2023-02/Americans_Cant_Consent.pdf
- UNESCO. (2022c). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137
- UN (United Nations). (1948). *Universal Declaration of Human Rights*. www.un.org/en/about-us/universal-declaration-of-human-rights
- UN. (1966). *International Covenant on Civil and Political Rights*. General Assembly Resolutoin 2200A (XXI). www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights
- UN. (2023a). *Information Integrity on Digital Platforms*. Our Common Agenda, Policy Brief 8. www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf
- UN. (2024b). *Pact for the Future, Global Digital Compact, and Declaration on Future Generations*. www.un.org/sites/un2.un.org/files/sotf-pact_for_the_future_adopted.pdf
- UN. (2024c). *Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development*. General Assembly A/78/L.49. https://digitallibrary.un.org/record/4043244/?v=pdf
- Urman, A., & Makhortykh, M. (2024). The silence of the LLMs: Cross-lingual analysis of political bias and false information prevalence in ChatGPT, Google Bard, and Bing Chat. OSFPreprints, pp. 1-11. doi:10.31219/osf.io/q9v8f.
- US State of California. (2018). *California Consumer Privacy Act (CCPA)*. State of California Department of Justice. https://oag.ca.gov/privacy/ccpa
- Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1), 1-13. https://doi.org/10.1177/2056305120903408
- van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197-208. https://doi.org/10.24908/ss.v12i2.4776
- van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, *1*(1), 2-14. https://doi.org/10.12924/mac2013.01010002
- van Dijck, J., de Waal, M., & Poell, T. (2018a). *The Platform Society: Public Values in a Connective World*. Oxford University Press.
- van Dijck, J., Poell, T., & de Waal, M. (2018b). News. In J. van Dijck, T. Poell, & M. de Waal, *The Platform Society: Public Values in a Connective World* (pp. 49-72). Oxford University Press.

- Vanoli, C. (2024). Moving AI governance from principles to practice. *ITU News*, 19 April. www.itu.int/hub/2024/04/moving-ai-governance-from-principles-to-practice
- Veale, M., Matus, K., & Gorwa, R. (2023). AI and global governance: Modalities, rationales, tensions. *Annual Review of Law and Social Science*, *19*, 255-275. https://doi.org/10.1146/annurev-lawsocsci-020223-040749
- Velkova, J. (2024). Dismantling public values, one data center at the time. NordMedia Network, 20 February. https://nordmedianetwork.org/latest/news/dismantling-public-values-one-data-center-at-the-time
- Verdegem, P. (2021). Tim Berners-Lee's plan to save the internet: Give us back control of our data. *The Conversation*, 5 February. https://theconversation.com/tim-berners-lees-plan-to-save-the-internet-give-us-back-control-of-our-data-154130
- Verhulst, S. G. (2023). Steering responsible AI: A case for algorithmic pluralism. ArXiv. https://doi.org/10.48550/arXiv.2311.12010
- WAN-IFRA (World Association of News Publishers). (2023). *Gauging Generative AI's Impact on Newsrooms: Survey: Newsroom Executives Share Their Experiences So Far*. World Association of News Publishers. https://wan-ifra.org/insight/gauging-generative-ais-impact-in-newsrooms
- Washington, J. (2023). *Combating Misinformation and Fake News: The Potential of AI and Media Literacy Education*. http://dx.doi.org/10.2139/ssrn.4580385
- Wasserman, H., & Madrid-Morales, D. (2019). An exploratory study of 'fake news' and media trust in Kenya, Nigeria and South Africa. *African Journalism Studies*, *40*(1), 107-123. https://doi.org/10.1080/23743670.2019.1627230
- WEF (World Economic Forum). (2024). How to manage AI's energy demand – today and in the future. 25 April. www.weforum.org/agenda/2024/04/how-to-manage-ais-energy-demand-today-tomorrow-and-in-the-future
- Weismueller, J., Gruner, R. L., Harrigan, P., Coussement, K., & Wang, S. (2023). Information sharing and political polarisation on social media: The role of falsehood and partisanship. *Information Systems Journal*, *Special Issue, 34*(3), 854-893. https://doi.org/10.1111/isj.12453
- Werthner, H., Ghezzi, C., Kramer, J., Nida-Rümelin, J., *et al*. (Eds) (2024). *Introduction to Digital Humanism: A Textbook*. Springer Nature Switzerland.
- Wiggers, K. (2019). IBM releases Diversity in Faces, a dataset of over 1 million annotations to help reduce facial recognition bias. VentureBeat, 29 January. https://venturebeat.com/ai/ibm-releases-diversity-in-faces-a-dataset-of-over-1-million-annotations-to-help-reduce-facial-recognition-bias
- Windwehr, S., & York, J. C. (2020). Facebook's most recent transparency report demonstrates the pitfalls of automated content moderation. Electronic Frontier Foundation, 8 October. www.eff.org/deeplinks/2020/10/facebooks-most-recent-transparency-report-demonstrates-pitfalls-automated-content
- WIPO (World Intellectual Property Organization). (2024). *Getting the Innovation Ecosystem Ready for AI: An IP Policy Toolkit*. Frontier Technologies Division. www.wipo.int/publications/en/details.jsp?id=4711
- Wu, J., Gan, W., Chen, Z., Wan, S., & Lin, H. (2023). AI-generated content (AIGC): A survey. *ArXiv*. https://arxiv.org/pdf/2304.06632
- Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., *et al.* (2022). Sustainable AI: Environmental implications, challenges and opportunities. ArXiv. https://doi.org/10.48550/arXiv.2111.00364
- Wulf, A. I. J., & Seizov, O. (2022). 'Please understand we cannot provide further information': Evaluating content and transparency of GDPR-mandated AI disclosures. *AI & SOCIETY*, *39*, 235-256. https://doi.org/10.1007/s00146-022-01424-z
- X. (2024). About Community Notes on X. X Help Center. https://help.x.com/en/using-x/community-notes
- Yan, B., Li, K., Xu, M., Dong, Y., *et al*. (2024). On protecting the data privacy of large language models (LLMs): A survey. *ArXiv*. https://doi.or/10.48550/arXiv.2403.05156
- York, J. C. (29 October 2022). Elon Musk doesn't know what it takes to make a digital town square. MIT Technology Review, 29 October. www.technologyreview.com/2022/10/29/1062417/elon-musk-twitter-takeover-global-democracy-activists
- Zhao, Y., & Chen, J. (2022). A survey on differential privacy for unstructured data content. *ACM Computing Surveys*, *54*(10s), 207, 1-28. https://doi.org/10.1145/3490237