INTERNET POLICY REVIEW
Journal on internet regulation

Volume 12 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Governing artificial intelligence in the media and communications sector

Jo Pierson *Hasselt University* jo.pierson@uhasselt.be
Aphra Kerr *Maynooth University* aphra.kerr@mu.ie
Stephen Cory Robinson *Linköping University*
Rosanna Fanni *Centre for European Policy Studies (CEPS)*
Valerie Eveline Steinkogler *Vrije Universiteit Brussel*
Stefania Milan *University of Amsterdam* s.milan@uva.nl
Giulia Zampedri *Vrije Universiteit Brussel*

DOI: https://doi.org/10.14763/2023.1.1683

Published: 21 February 2023
Received: 14 December 2021 **Accepted:** 3 May 2022

Funding: Jo Pierson received support from the DELICIOS project (Delegation of Decision-Making to Autonomous Agents in Socio-Technical System) funded by The Research Foundation – Flanders (FWO) (Grant G054919N) and Aphra Kerr would like to acknowledge support from the ADAPT Centre which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106_P2).
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Pierson, J. & Kerr, A. & Robinson, S. C. & Fanni, R. & Steinkogler, V. E. & Milan, S. & Zampedri, G. (2023). Governing artificial intelligence in the media and communications sector. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1683

Keywords: Artificial intelligence, Trust, AI governance, European policymaking

Abstract: The article analyses critical blindspots in current European Artificial Intelligence (AI) policies and examines the potential impact of data and AI in the emerging socio-technical ecosystem of the contemporary Media and Communications (MC) sector from the perspective of critical media and communication studies. We first identify central blind spots in the dominant EU trustworthy and risk-based approach to governing AI. Next, we propose a novel multi-level framework to analyse key policy challenges for governing AI in the MC sector. The framework and discussion are based on desk research and multi-stakeholder expert discussions. The article concludes with reflections on AI governance in development, deployment and use in the MC sector.

Introduction

Since 2018 several European policy efforts have aimed to incorporate ethics principles and European fundamental rights into the governance of artificial intelligence (AI) (Floridi et al., 2018; Hagendorff, 2020). The European Council (2017) announced its priority to establish a "high level of data protection, digital rights and ethical standards" for AI systems (p. 2). The ensuing Communication from the European Commission (2018) claims to address the ethical and legal questions surrounding AI, which is reiterated in the 2019-2024 Agenda for Europe (von der Leyen, 2019). The subsequent multistakeholder deliberations through the High Level Expert Group on Artificial Intelligence (AI HLEG), including the AI Ethics Guidelines and the subsequent European Commission White Paper on AI and the AI Alliance multistakeholder efforts, all aim to render broad abstract principles actionable. This includes the Assessment List for Trustworthy AI (ALTAI) developed by the AI HLEG, a checklist for AI developers and deployers to use when implementing AI ethics principles in their operations (HLEG, 2020).

Notwithstanding this "principle proliferation" (Floridi & Cowls, 2019), such initiatives generally fall short in addressing the broadly defined impact of AI technologies on the media and communication (MC) sector. Neither existing policy initiatives nor the literature fully consider how AI affects this sector, limiting their focus most notably to automated content moderation on social media or news sites, or the distribution of disinformation across social networks. This paper fills this gap by exploring the interplay between AI technologies and the MC sector. This sector encompasses a variety of contemporary forms of technologically supported content and communication. From the vantage point of critical media studies and critical social science, we offer two key contributions. Firstly, we identify key blind spots in the dominant trustworthy and risk-based approach to AI being pursued by the European Union. These blindspots include a lack of attention in AI policies to the potential impact of AI on the infrastructural and structural asymmetries in the MC sector, a tendency to ignore the cultural, social and democratic importance of diverse media to societies and a failure to attend to the ways in which AI can be used to undermine social cohesion and collective rights. Our contribution exclusively assesses the key non-regulatory policy documents up until Spring 2021, due to their primary relevance in establishing the "Proposal for a regulation on a European approach for AI" (AI Act) in April 2021 (European Commission, 2021). Secondly, we propose a multi-level approach to analysing the key policy challenges to governing AI. Such a perspective is needed for an understanding of the sectoral specificities and the socio-technical consequences of AI in the MC sector, more

generally. A multi-level perspective is needed to understand the key areas in which AI is currently applied in this sector. To this end, we identify four dominant uses of AI in the MC sector: automating data capture and processing, automating content generation, automating content mediation and automating communication. The four levels are the outcome of extensive desk research and our expert committee meetings, part of a much larger European multi-stakeholder initiative established by the Atomium institute in Brussels. For each of these uses, we identify overarching opportunities and risks stemming from the use of AI in relation to the AI HLEG requirements for trustworthy AI. In doing so, we propose a novel analytical and macro-level framework to identify the key issues raised by the application of AI in the MC sector. We lay the groundwork for identifying who should be responsible for addressing these issues and how they might do so.

The article comprises five sections. After the introduction, Section 2 delineates how the MC sector and related AI implementations are conceptualised in the article. Section 3 summarises and assesses state-of-the-art European AI policy initiatives and governance, offering a critical perspective. Section 4 analyses the MC sector and through case studies shows how the 7 key requirements, as suggested by the AI HLEG, operate across four different levels. Section 5 offers recommendations to ensure that AI development and adoption is compliant with the key requirements for trustworthy AI in the MC sector, while reflecting on the adequacy of the 7 key requirements and the EU risk-based regulatory approach from the perspective of media and communications scholarship. We conclude by recommending three areas of oversight for governing AI: addressing data power asymmetry, empowerment by design for mitigating risks, and cooperative responsibility through stakeholder engagement.

Understanding AI and the media and communications sector

Defining and setting the boundaries of the MC sector is no simple matter given the fast evolution of digital technologies, corporate mergers and new market entrants. MC includes multiple forms of technologically supported content, interaction and communication. It embraces digital media, i.e. digitised legacy media, as well as born-digital platforms. The latter act as socio-technical intermediaries that collect and process data to enable and steer communication (Helberger, Pierson & Poell, 2018). Mobile applications (apps) are increasingly central to MC businesses, as much digital media communication today takes place in the context of an app. These apps are made available within a (mostly) privatised platform and communi-

cations infrastructure that is optimised to exploit data (Pierson, 2021; Nieborg & Helmond, 2019). Many forms of digital media increasingly rely on the data analysis of users and content. Data is collected, processed and evaluated in the MC sector for many purposes, including the automated personalisation of content (e.g. news recommendations) and targeted advertising. Finally, the media content and communication sector rely upon existing telecommunications and internet infrastructure. The latter includes wired and wireless mobile internet and communication service providers. These operators and their services are a crucial part of the MC sector, and constitute what Winseck (2019, p. 176; 2020, p. 1) calls the "network media economy". From our perspective this sector includes those media industries usually included in the definition of creative and cultural industries, but also telecommunication service providers and parts of the wider computing industry, which provide infrastructure for the circulation of user data, media content and communication.

Our understanding of AI is based on the AI HLEG (2019a), which defines AI as human designed systems implemented in a digital or physical environment in the form of software-based systems or hardware devices. AI systems process and assess data based on reasoned decision-making, producing outputs and suggesting relevant actions to achieve given goals. In other words, AI systems intervene in society on the basis of the data they collect, and decisions the software is enabled to make. This process is guided by a set of symbolic rules or a numeric model, and also by AI's ability to learn from its environment and previous outputs (HLEG, 2019a).

Given that AI systems provide information about the world, make recommendations that may impact our everyday decisions, and provide value-based judgements, citizens must be able to trust them (Ferrario et al., 2019). According to the AI HLEG, trustworthiness should represent a "prerequisite for people and society to develop, deploy and use AI" (HLEG, 2019b). This trust has been severely impacted by public revelations about the use of AI by states and companies for the purpose of public surveillance and manipulation (Kerr et al. 2020). Research has also revealed potential and actual harms, biases and discrimination (see, e.g., Hintz et al., 2019; Eubanks, 2018). These developments and findings suggest that the MC sector needs to be vigilant to ensure that AI systems and data processing are, and remain, trustworthy.

The Charter of Fundamental Rights of the European Union provides a sound basis upon which to build policies that protect citizens from being unwillingly influenced by AI and having their personal data used without their consent (European

Parliament, 2000). The Charter emphasises the importance of freedom of expression, access to information and the protection of personal data, all essential aspects of democracy and the public sphere, but all are undergoing transformation in the current digitalisation and datafication phase of MC. Special consideration is also given to children, minority groups such as migrant communities and marginalised communities including LGBTQIA+ and refugees. Therefore, we interpret "trustworthy AI" as AI systems that are worthy of the European public's trust from a fundamental rights perspective. We focus on the European context, given the EU's ongoing commitment to establish a regulatory and societal framework for "trustworthy" and "human-centric" AI. The next section briefly examines the evolving EU regulatory regime and identifies some critical blindspots in relation to AI applications in the MC sector.

Critical blindspots in the European perspective on AI governance

An emphasis on democratic values, fundamental rights and the rule of law is integral to the EU vision for the development, deployment and governance of AI. At the same time, AI is a European research and innovation priority. When European principles and values are translated into emerging policies, which are also meant to support innovation, we see conflicting goals and varying emphasis to governing AI between EU institutions and national governments. The framing of emerging regulatory approaches is especially relevant for the MC sector, where automated decision making and content generation technologies are increasingly prevalent. Consider, for example, the use of AI across the entire news production process: automated content production (Willens, 2019), machine translation of text (ADAPT, 2021), correction and transcription services (Marr, 2018), facts and potential fake news (Cassauwers, 2019) or data-driven tools to personalise users' news feeds and recommend other content (Warren, 2021).

The AI HLEG, appointed by the European Commission (EC) in June 2018, consisted of 52 members and included companies and industry representative groups, academics from computer science, law, philosophy and economics, and representatives of civil society groups and digital rights organisations. Notably, no academics from media and communication studies, nor people from traditional public service media or new digitally native European media companies were represented, whereas a number of telecommunications and technology companies, including Alphabet-Google, were. Neither the draft nor final report of the AI HLEG mentioned media specifically, although both mention that policy and regulation needs to pay atten-

tion to "situations with asymmetries of power or information" (HLEG, 2019b: 13). There was also no discussion of AI-driven mis- or disinformation.

The final AI HLEG report, "Ethics Guidelines for Trustworthy AI", (HLEG, 2019b) listed seven requirements for trustworthy AI: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination, and fairness, 6) societal and environmental wellbeing, and 7) accountability. The AI HLEG final report was influential in shaping a "European" approach to governing AI in that it further substantiated the concepts of "ethical" and "trustworthy" AI (Hasselbalch, 2020), and was foundational for the Assessment List of Trustworthy Artificial Intelligence (ALTAI) (HLEG, 2020) and the February 2020 White Paper on Artificial Intelligence (EC, 2020). In the White Paper, the EC highlights the use and potential impact of AI (1) for information selection and content moderation by online intermediaries; (2) in tracing people's daily habits; and (3) in creating information asymmetries by which citizens might be left powerless. The White Paper considers AI systems "high-risk" if "both the sector and the intended use involve significant risks" (EC, 2020), particularly if safety, consumer rights or fundamental rights are at stake. It is noteworthy that the White Paper sees no specific "high-risk" issues, nor mentions any use examples from the media, creative or cultural industries – while health, security, energy, farming, transport etc. are all mentioned. While the high-risk distinction promotes a risk-based approach to AI, incremental, long-term and societal-collective risks are not addressed. The White Paper, furthermore, focuses on AI as a tool or technology which disregards larger power imbalances and power structures that derive from the use of AI in a larger digital communication infrastructure.

Subsequently, the EC published a proposal for the regulation of AI in April 2021. This "AI Act" (AIA) adopts the risk-based regulatory approach. Certain applications of AI, including AI systems that could cause physical harm (MacCarthy & Propp, 2021), would be banned under the AIA, and high- and limited-risk AI systems (European Commission, 2021) would be regulated. High-risk AI systems would only be put on the market if they have conducted conformity assessments in advance and meet obligations such as transparency, human oversight, record-keeping, robustness, accuracy and security (European Commission, 2021). The proposed regulation focuses mainly on AI system providers and relies on companies and organisations to self-assess, and on national supervisory authorities to oversee compliance with the regulations (MacCarthy & Propp, 2021). The AIA does not include specific consumer rights, such as the right to redress.[1] However, consumers must be informed

1. See Fanni et al. (2022) for more extensive discussion on redress in AI systems.

when they are interacting with an AI system, or when their emotions are being detected (for the purpose of providing relevant services, products, etc.).

To summarise, the European policy documents we have analysed emphasise generating the conditions for trustworthiness in the development, deployment and use of AI, but fail (to date) to consider the long-term, incremental risks for democratic processes and fundamental rights. The AIA draft currently fails to provide sufficient guidance on accountability or redress mechanisms to address harms. In these documents there is a lack of attention paid to the MC sector, even though there is a call for sector-specific analysis of threats and risks. A healthy democracy depends on independent and diverse news and media outlets, especially in their role as the Fourth Estate (Csaky, 2021). However a growing part of our society, up to two thirds (Newman et al., 2021), consumes news, and consequently, forms their opinions from sources on social media. Developments like AI-driven recommender systems in digital media have helped spread online disinformation, and have been implicated in the radicalisation and polarisation of societies, partly due to their advertising-driven preference for arousing emotional and controversial content (Coalition to Fight Digital Deception, 2021; Hagey & Horwitz, 2021; Ribeiro et al., 2020). These are documented risks which are detrimental to democracy and social cohesion. It is not clear, however, if the AIA would classify these applications as high or medium risk.

From the MC sectoral perspective, our analysis of European policies related to AI governance thus reveals several fundamental blind spots. In general the focus on algorithms, data and information in AI policy documents prioritises transmission and access issues. This approach treats data and information as bits to be transferred rather than being concerned with the diversity, veracity or meaning of the content that is being generated, issues which have long been at the heart of media policy and regulation in many European countries. Indeed, we concur with the critique that the European perspective on AI regulation is technologically determinist and lacks the attention to redress relative market powers (Veale, 2020). For us the key blindspots are the lack of attention to: 1) the infrastructures of surveillance capitalism and datafication which prioritise data gathering and engagement (Zuboff, 2019), 2) a trend towards increasing corporate and market concentration and power asymmetries in MC sectors (often by large companies based and regulated outside Europe), 3) a tendency to ignore the important cultural, social and democratic role of the media, including the importance in the diversity of content and representation (Napoli, 2019), 4) a lack of understanding of how AI is used to categorise people in new ways, and to target and adapt content and communica-

tions towards them in opaque ways, and finally, 5) a preference from social and digitally native media and technology companies for self-governance and minimal content or communication oversight. Without due consideration to these blindspots a generic commitment to trustworthy AI principles and a reliance on corporate self-certification is not, in our opinion, sufficient. In short, most European AI policies neglect relevant issues that undermine trust in AI from an MC perspective or – at best – label them as low-risk.

Background and methodology

This paper emerged following the involvement of the authors in the AI4People programme run by the Atomium European Institute for Science, Media and Democracy (Atomium-EISMD 2018-2020). The Atomium initiative aimed to develop 7 industry-specific frameworks and policy recommendations for governing AI technology. Given the lack of EU AI MC sector-specific policy proposals as identified above, the authors of this paper welcomed the chance for intersectoral dialogue in the context of the Atomium initiative. They were invited to join a 12 member committee of the AI4People programme focused on the MC sector, appointed by Atomium-EISMD in consultation with chair Jo Pierson, including equal numbers of MC academics (6) and personnel from media and technology industry associations and multinational companies (6).[2]

The starting point of the Committee's work was the "Key Requirements for Trustworthy AI", as proposed by the European Commission (EC) High-Level Expert Group on Artificial Intelligence, henceforth AI HLEG (HLEG, 2019b). The goal was to discuss and debate how to translate the AI HLEG's 7 key requirements for Trustworthy AI into a framework that could be implemented in the MC sector, and to identify opportunities and risks specific to the sector. This committee produced a 30-page report in December 2020, which is available online (Pierson et al., 2021). This article brings together the academics from the committee to reflect further on the AI4People Media and Communication committee discussions, (March–November 2020), and to update our analysis to take account of subsequent policies and draft regulations which emerged during the first half of 2021.

The four levels of intervention of AI in the MC sector described in the next section were identified through both desk-based research and discussion, as well as debate and analysis by the MC committee members, based on a multistakeholder ap-

2. Full membership information retrieved July 7, 2022, from https://www.eismd.eu/ai4people/committees/committee-on-media-and-technology

proach (Raymond and DeNardis, 2015). First, we conducted extensive desk research on existing frameworks at the interface of AI, media, communication and EU policy. Next, a total of 44 case studies containing best practice and problematic practice use cases of AI in the MC sector were solicited from committee members and our wider networks based on their expertise and practical experience. Additional information, literature and explanation on the cases was provided. The case study repository was examined and overlapping thematic issues were identified. In a final step, the AI case studies were grouped into four themes, cross-combined with the 7 key requirements and used as the basis to develop our multi-level approach to evaluate trustworthy AI in the MC sector, and to assess the adequacy of current EU AI policies to govern the use of AI in the MC sector.

A multi-level analysis of trustworthy AI in media and communication

In order to evaluate the current European approach to trustworthy AI, from the MC sector perspective, we developed a multilevel framework of current AI applications. These were:

- Level 1: Automating data capture and processing
- Level 2: Automating content generation
- Level 3: Automating content mediation
- Level 4: Automating communication

The four levels broadly correspond to different stages in the typical (big) data life cycle (Jagadish et al., 2014).[3] The four MC AI levels are not mutually exclusive, and in what follows we illustrate our discussion at each level with concrete examples from our case study research. This approach allows us to identify tensions between the 7 key requirements, while also considering complex issues that occur predominantly on one level (e.g. advertising in data capture and processing), but may also occur at other levels. We also identify high-risk applications in the MC sector which should be considered in EU AI policy.

Level 1: Automating data capture and processing

The first level encompasses a variety of AI technologies that systematically capture and process data in the MC sector. This typically includes data capture and processing by digital media, platforms and websites for profiling, personalisation,

3. Stages for creating value from (big) data in a multi-step process: acquisition, information extraction and cleaning, data integration, modelling and analysis, and interpretation and deployment.

inferential predictive analytics, targeted advertising, etc. It also includes facial and voice recognition systems capturing emotional expressions, as well as GPS/location tracking and VR/AR.

The principles of human *agency and oversight* and *privacy and data governance* are most pertinent at this level. First, the EU legislative framework, in particular the General Data Protection Regulation (European Parliament & Council, 2016) and consumer protection standards, already protect individual rights to allow citizens to make informed and independent choices. These protections also apply to automated data capture and processing in AI systems: citizens should always be able to decide if and how they choose to use a certain service or be tracked by it. European citizens using a MC service should have the right to decide what and how much data will be collected, what it will be used for, where it originates from and how it will be shared. Despite the existence of these legal protections and ethical principles, many AI-driven applications and services rely on an advertising and a data-driven business model that remains opaque. This is particularly relevant for online behavioural advertising, where internet users' behavioural data (website visits, clicks, mouse movements, etc.) and metadata (browser type, location, IP address, etc.) are collected and processed to create profiles that will be used to personalise ads and improve conversion rates.

Civil society groups have argued that automated advertising systems with real-time bidding (RTB) have been capturing and processing data in prohibited and unethical ways (Irish Council for Civil Liberties, 2021; Information Commissioner's Office, 2019). RTB is the system by which advertisers bid on the possibility of instant targeted advertising to website visitors by using personal data that is collected through tracking and is shared with all bidders. Even advertisers who do not win the auction receive personal data that ascertains the visitor's interest at the conclusion of the auction. Some advertisers allegedly participate in the auctions merely to enrich their data sets. The targeting is based on profiles of users built via extensive and persistent tracking of online and (possibly) offline activities (e.g. via cookies or pixels). Users' past behaviour is a category of these profiles, but so are inferred preferences and affinities, often including sensitive categories protected by the GDPR. For example, Alphabet-Google and several data brokers have been accused of violating EU data protection rules by harvesting and processing people's personal data to build detailed online profiles, including information on sexual orientation, health status and religious beliefs (Scott & Manancourt, 2020), and in turn grouping people based on their assumed interests rather than their personal traits, becoming a digital construction of groups based on categories of

identity (Cheney-Lipold, 2011).

In addition, given the value that can be generated by collecting and processing extensive amounts of (personal) data and using it to personalise content and advertising, special attention is needed to safeguard a level playing field between large and small operators to ensure the diversity of operators and content. According to the GDPR (Recital 71), personalisation through profiling involves the automated processing of personal data to evaluate individuals' personal aspects, to analyse or predict their economic situation, health, preferences or behaviour. The GDPR (Art. 22) grants individuals the right not to be subjected to automated decision-making if it has a significant effect on their lives. When automated profiling and personalisation have significant effects, it will only be allowed if users consent. Although all companies and organisations in the media and advertising ecosystem who process European citizen data are subject to the GDPR, larger operators with more financial resources have made challenging regulatory judgements under the GDPR since it was introduced. Regardless, they are also better able to pay the fines imposed. Smaller media companies and community media with fewer resources are not only losing advertising revenue to a small number of multinational platform companies, they are also unable to compete on technical grounds. Ultimately, this model may result in less diverse and less local content generation. The consolidation of personal data in fewer hands might also increase, and perversely, negatively affect people's rights and freedoms overall, but also their cultural and communication rights, and media diversity, per se (Kerr et al., 2019). An interesting attempt to counter this consolidation of power is the development of pre-commercial data pooling and communal processing initiatives to benefit several (competitive) companies at once in local markets (e.g. Ads & Data in Belgium) (van Zeeland et al., 2019).

To summarise, the use of AI to automate data capture and processing reinforces problems of operator concentration and power in the sector, and further threatens cultural and communication diversity, as well as rights. The application of existing legislation, such as the GDPR, and existing trustworthy AI principles, have thus far failed to address these issues. The speed, scale and opaqueness of new AI methods, and the complexity of data and advertising infrastructures, are undermining the ability of end users and media companies to give and receive informed consent.

Level 2: Automating content generation

The second level refers to online content produced by automated systems, either

entirely or in combination with human agents. Examples of common AI uses in content generation are news reporting apps (based on user preferences),[4] translation tools, and – critical from a democratic perspective – disinformation and deep fakes: images or videos in which identities and/or elements are digitally manipulated. Such AI-generated content may appear in text, image, audio or video. How much content is generated automatically and to what extent users are informed of this fact is unclear.

Journalists, and the companies and organisations they work for, play a major role in providing trustworthy news and information. Thus, *human agency and oversight, transparency, accountability* and *societal well-being* are highly relevant to automated content generation. In data journalism, for instance, AI helps to identify patterns in large datasets. AI-driven tools can suggest titles and photos, help writers find a new angle to a topic, and produce draft versions of articles. Automated systems assist the journalist in writing the story, but the journalist should remain the primary storyteller (Willens, 2019). Assistive content generation with human editorial input and *human oversight* allows for faster content generation. The news generation in some news genres, especially fact-based ones, is highly automated. For instance, specialised natural language processing tools can generate sports articles and financial reporting (Peiser, 2019); recent projects even involve video reporting (Chandler, 2020).

The use of AI at this level has had a significant impact. The increasing pace and efficiency of automated news production can put pressure on smaller newsrooms, which usually do not have access to large datasets and robust AI-systems (Helberger et al., 2019). Higher automation of content generation can lead to job losses, impacting the diversity of journalists, as well as the bias, accuracy and diversity of content and wider *societal well-being* (Srnicek, 2017; Lindén & Tuulonen, 2019). Automating content generation will also create new kinds of jobs, which may require substantial upskilling, training and education. In addition, content produced by AI systems is often not flagged as such to the user, undermining the principle of *transparency*. Since trust in news and information is often associated with its producers, great care needs to be taken with the use of AI tools in content production, as this can reduce the trustworthiness of content overall.

Technical robustness is also highly relevant in producing and translating texts. Machine translation risks replication biases (e.g. stereotypes, gender and racial biases) and errors from training datasets, affecting the principle of *diversity, non-dis-*

4. For example, Google News, Apple News, Reddit, Digg and Flipboard.

crimination and fairness. Robust machine translation, in contrast, can help to preserve the EU's linguistic and cultural plurality. For example, the ADAPT research centre (2021) in Ireland develops datasets and intelligent models that automatically translate online content for native speakers of low-resource languages, and make important content available to people in their language of choice. Projects have focused on developing resources for Irish, Serbian, Basque and non-European languages, including Hindi. Their approach employs both AI and humans, rather than fully automated systems.[5]

On social media platforms and in messaging services, deep fakes generated by AI-driven tools are increasingly common. These formats simulate in an increasingly realistic manner a speech or an action, usually by a public persona (such as politicians, celebrities and actors). Such false information is often generated without the individual's knowledge, and viewers may be unaware that the video was altered. Thus the principles of *human agency* and *societal wellbeing* are at stake. Deep fakes can foster the spread of contentious and harmful content like "fake news", disinformation and hate speech. A recent study found that 72 percent of people reading an AI-generated news story thought it was credible (Leibowicz, 2019). Generating deep fakes and producing disinformation challenges media integrity, erodes trust and has negative implications for democracy. This relates to the two main roles of media in a healthy democracy: correctly informing citizens to support meaningful political choices ("watchdog function") and creating a diverse public forum that allows different ideas and opinions to be shared and discussed (Helberger, 2019; Balkin, 2018). Deep fakes and disinformation also impact *diversity, non-discrimination and fairness*. Current forms of redress to tackle these issues seem to be inadequate (Fanni et al., 2022). Deep fakes are specifically mentioned in the EC draft regulatory framework on AI, but without specifics on redress (European Commission, 2021). Most recently, the EC introduced a new redress and liability regime applicable to AI systems causing damage, which is a first step towards meaningful redress rights for individuals. The EU AI Liability Directive (European Commission, 2022) strengthens users' rights to access to information and alleviates victims' burden of proof in the case of harm by the fault or omission of an AI provider, developer or user.

5. For example, they have developed a high-quality Irish-English system called Tapadóir to translate documents into Irish for the Irish government. From 2021 on, all European documents will also have to be translated into Irish and much of this will be done using these automated systems supplemented by Irish language native speakers and translators.

Level 3: Automating content mediation

The third level involves automated filtering systems in the distribution and moderation of online content and advertising. AI technologies in content distribution include recommender systems for entertainment and social media content, online news aggregators, and programmatic advertising (including RTB). These provide user-specific and context-conforming content. Other AI systems moderate content to detect and tackle contentious content like fake news, mis- and disinformation (European Parliament. Directorate General for Parliamentary Research Services, 2019a) and harmful content (Lacoma, 2020).

Employing automated filtering systems in online content and advertising mediation tasks entails the principles of *diversity, non-discrimination, fairness, human agency and oversight*. For example, years after the initial research into discrimination in online employment ads, higher salary positions are still advertised to predominantly (assumed) male users (Burke, Sonboli & Ordonez-Gauger, 2018). When AI technologies act as gatekeepers and set agendas in the online sphere, they can co-determine what people see or do not see, as well as what content users can generate online. This could affect freedom of expression, media diversity and the plurality of voices (Helberger et al., 2018). Algorithmic content distribution can constrain access to diverse information and create "filter bubbles" leading to "echo chambers", i.e. personalised content. Online platform recommender systems tend to magnify hyperactive users' interests and content, or certain producers, while passive users' interests and other forms of content become invisible (Content Personalisation Network, 2020; Papakyriakopoulos et al., 2020). However caution is needed to not overestimate the social impact of these algorithmic developments (Löblich & Venema, 2021; Möller et al., 2018). Hence, political microtargeting and opinion formation could become subject to (un)intentional algorithmic manipulation and media content could become less diverse (Feezell, Wagner & Conroy, 2021).

Algorithmic filtering systems are required given the high volume and fast-paced production of online content. AI systems are crucial assistants to humans in the evaluation of harmful content such as child abuse, racism and harassment. These tools may reduce the physical and mental impact of this work on the human moderators (Ofcom, 2019) by flagging harmful content, blurring particularly harmful sections or engaging in "visual question answering". AI can also be used to tackle malicious online behaviour directly using notifications or chatbots that make the user aware that a post contains harmful content, or the technology can create a short delay in the posting process to encourage the user to rethink the message

(Statt, 2020). AI systems can also provide alternative content suggestions that are more positive, but express the original message. In both instances, the human agent – a content moderator or user – ultimately decides. However, AI systems have also been found to have many limitations when it comes to understanding the complexity and diversity of fast changing communications (Kerr et al., 2020; Gowra et al., 2020).

Transparency and *accountability* are two major principles for tackling algorithmic decision-making and countering potential abuse. The highly complex architecture of AI moderating tools, and their proprietary nature, makes it difficult to understand and assess their decision-making process (European Parliament. Directorate General for Parliamentary Research Services, 2019b). Inappropriate algorithmic standards, combined with a lack of resources, can result in negative content being left online, and/or appropriate content being removed. Transparency is the first step to developing systems of accountability for algorithmic decisions and judgements, making these processes legible to different stakeholders, e.g. general public, certain sectors, human agencies and/or oversight bodies. Abuse of algorithmic filtering systems could result in censorship, which would violate democratic principles. To prevent this, civil society groups have made software for algorithmic auditing methods open source to expose personalisation algorithms on social media and shopping platforms, for example Algorithms Exposed (ALEX) (see Beraldo et al., 2021 and Milan & Agosti, 2019). Their goal is to empower both advanced users and low-skill users with data extraction tools, and to enhance individual and societal knowledge of algorithmic content mediation.

Level 4: Automating communication

The fourth level in our analysis includes all forms of interaction and communicative actions and infrastructure enabled by AI, including: speech and language technologies, chatbots, smart speakers, voice assistants and automated marketing communication. All these AI systems that simulate a realistic conversation, and the encoding and decoding of conversational messages and data from users fall into this level.

The most important requirements for automated communication are *human agency and oversight, diversity, non-discrimination, fairness* and *transparency*. First, *transparency* would entail all AI-empowered communication channels being open, or making much of their data infrastructure auditable, including how information and output is compiled. *Transparency* would also enable users to better understand how their conversation data is used and evaluated. Especially in automated mar-

keting and communication, users can fall prey to misleading messages or biased information, which could be avoided if *transparency* in marketing practices were mandatory.

Open curated datasets may improve the experience of AI systems users in terms of *diversity, non-discrimination and fairness,* and prevent the distortion of conversation and information between humans and machines. Further, considering the European linguistic diversity, automated communication systems can already discriminate against or disadvantage certain linguistic minorities. Finally, users should be able to choose whether they want to interact with a chatbot or with a human being, as reflected in the principle of *human agency and oversight*. This also links to the principle of *accountability* if a consumer-oriented chatbot, e.g. for a bank, gives imprecise or incorrect information that can cause harm or damage.[6] Accountability in automated communication must therefore include redressing automated decisions by chatbots. The AIA refers to ensuring accountability mechanisms for affected persons through transparency and traceability, as well as ex-post controls (European Commission, 2021). This fits in the computer science tradition of explainable AI (XAI). XAI is defined as the practice of improving understandability, trustability, and the manageability of emerging AI systems (Meacham et al., 2019), however all too often the focus is on making these systems understandable to AI experts rather than end users.

The *societal and environmental wellbeing* principle is crucial to overall decisions about whether it is suitable, viable, sensible and sustainable to adapt AI enabled automated communication for certain cases. Automated communication AI tools require extensive datasets and the training of resource intensive models in order to automatically translate online content. While machine translation may increase social inclusion for speakers of low-resource languages (ADAPT, 2021), thereby making important content accessible to linguistically diverse communities, these models also risk replicating biases and errors from training datasets. The fact that employment opportunities for translators are significantly diminished by automated communication AI technologies threatens the *societal wellbeing* principle, which means that automated communication AI companies were required to mitigate the impact of their technologies on the traditional job sector. In addition, an increasing concern is the environmental costs of training large AI models, running statistical analysis and cooling data infrastructures. This is a hidden societal cost at all levels that deserves critical attention (Crawford, 2021).

6. However, this also applies to wrong information from a human employee, and in both cases the bank is liable anyway.

Towards trustworthy AI in media and communication

This section spells out some key considerations for addressing the critical blindspots in current AI policy in Europe, moving towards the implementation of trustworthy AI in MC across the four levels discussed above. Three considerations are proposed: addressing data power asymmetries, empowerment by design for mitigating risks, and ensuring cooperative responsibility through diverse stakeholder engagement.

Addressing data power asymmetries

The use of AI is eroding meaningful, intentional and informed consent in the *automatic data capture and processing* of personal data by the MC sector; it is also deepening power asymmetries between MC organisations and their audiences/users. Do citizens and customers know when their personal data is being collected by AI-enabled systems? Users should be informed when their volunteered, observed or inferred personal data is being used to train machine learning algorithms before they decide whether to opt in. AI-driven businesses could be obliged to provide explanations on an ongoing basis, as has been suggested by Article 29 Working Party (WP29) in their guidelines on valid consent, which were recently updated by the European Data Protection Board (EDPB, 2020). Such positive data obligations enable citizens to act with *agency* in the face of data power (Kennedy et al., 2015). Clear and informed opt-in consent, combined with transparency obligations for algorithmic training and testing with user data, needs to be mandated for the MC sector. One approach might be to more widely embrace algorithmic registries, which detail the datasets that models were trained on and how algorithms are utilised, as Amsterdam and Helsinki have done (Moltzau, 2020).

Individual consent decisions will not prevent every harm stemming from abuses of automated personal data processing in the MC sector. While individuals may consent to the use of information about their emotions, political affiliation, health or sexual orientation, this use may have large-scale repercussions, for which individual choices cannot bear responsibility. Political microtargeting offers an example: individual users may consent to the use of data about their political preferences and emotional states on a platform, but aggregated data on attitudes and emotions linked to political preferences may be used to automatically manipulate other citizens' voting behaviours with potentially major societal effects, as the Cambridge Analytica scandal has illustrated (The Guardian, n.d.). Prevention of such malignant applications of automated data processing cannot rest on an individual's shoulders, but should be addressed with national or European regulation

based on an interdisciplinary, multi-stakeholder engagement of fundamental rights and public values. We recommend multi-stakeholder processes of value-sensitive design to investigate how predictive analytics, sentiment analysis and emotional AI threaten the integrity and autonomy of digital media users, especially in online behavioural advertising and synthetic content production. This approach is in line with the remit of Art. 22 GDPR ("Automated individual decision-making, including profiling").

Explainability is a complex, nuanced problem, considering the variety of European citizens and the complexity of some AI systems. Research and funding to increase AI *transparency* and explainability should be pursued and prioritised. However questions remain as to the adequacy of current technical solutions, as well as the fact that current XAI efforts are largely driven and funded by military organisations (e.g. DARPA) (Taddeo et al., 2021; Whittaker, 2021). Technical solutions should be supplemented with (co)regulatory efforts for establishing more transparency from digital platforms vis-à-vis independent regulators, especially on matters like internal processes for handling harmful and illegal content through algorithms and AI. Then we can better regulate platform-specific architectural amplifiers of contentious content, e.g. in recommendation engines, search engine features (such as autocomplete), features like "trending" and other mechanisms that predict what we want to see next. This approach fits with ex-ante principles-based co-regulatory approaches, when authorities attempt to force digital media companies to be more proactive in achieving state-determined public policy objectives. In that way, self-regulatory efforts can be better enforced, while avoiding purely punitive measures that penalise unlawful behaviour only after harm has been done (Vermeulen, 2019). Hence, we recommend strengthening research on process-based regulation and oversight on AI transparency and explainability, especially with regard to architectural elements for algorithmic amplification, e.g. avoiding widespread calls for violence by prohibiting algorithms from favouring and amplifying sensational news. Anticipatory data management policy should be a future priority in EU legislation. *Privacy and data governance* are moving targets, and new categories of personal data will be utilised, collected and created. Therefore, it is imperative to consider emerging sensitive AI-related personal identifiers, whether emotional data or even predictive AI systems, in the implementation and enforcement of GDPR, ePrivacy regulation revision and other European data-related policy initiatives (e.g. Data Governance Act, Digital Services Act, Digital Markets Act).

Empowerment by design for mitigating risks

Using AI technologies for MC activities like profiling, content and advertising per-

sonalisation threatens *human agency, transparency and safety*. Comprehensive solutions must be investigated and developed to address these threats. This fits in with the idea of "empowerment by design", i.e. building infrastructures and systems to give (organised) citizens (e.g. civil society organisations and activists) the agency to safeguard and strengthen their fundamental rights and the public interest (Pierson & Milan, 2017).

The advertising industry in the MC sector is a complex and multi-sided market with a multitude of actors, many of them intermediaries, such as networks of third parties with tracking technology, intermediary data brokers and exchanges, all competing in the market of RTB and automated auctions (Binns et al., 2018). Sensitive information about individuals – ethnicity, gender, sexual orientation, religious beliefs – can be inferred and used for online behavioural advertising and affinity profiling, i.e. grouping people according to their assumed interests rather than their personal traits, and through soft biopolitics: how biopower constitutes a population, and how that population is diffusely developed (Cheney-Lipold, 2011). Several scholars and digital rights organisations have made suggestions for empowering consumers against the illegal or unethical automated capturing and processing of their personal data (BEUC, 2020, pp. 16-17).

AI is also used in emotion detection in the MC sector. All these uses risk manipulating human behaviour and applying biased models and data from one context to another. These systems could "nudge" people into taking certain behavioural actions (Mele et al., 2021); infer belief and attitude; and incentivise use or concealment of certain emotional expressions. Emotion detection could likewise exacerbate existing biases against vulnerable groups. A set of actions could help to mitigate the risks posed by emotion detection AI. First, users should have to opt-in if any of their data is being used to detect emotions. User consent should be mandatory for MC businesses, as required by EU data protection law, with few exceptions. However, consenting to data collection is not sufficient; the issue is its application at scale. Those developing emotion detection AI need to facilitate full transparency and evaluation by relevant experts such as sociologists, psychologists, anthropologists, media scholars and psychiatrists. Overall, emotion detection in the MC sector is a high-risk application, and next to being listed as such in forthcoming regulations, a range of mitigating measures will need to be developed.

Cooperative responsibility and accountability through stakeholder engagement

Many concerns that arise in this sector can only be tackled by means and resources

beyond the sector. One way to secure *societal and environmental well-being* is to develop a shared responsibility for governance between civil society (users, human rights and consumer groups), public and private companies (platforms, technology and content producers, advertising and data service providers) and governments (education, policy and regulation). This type of "cooperative responsibility" requires digital media platforms, policy makers, users and other possible actors to develop a division of labour for managing their responsibility in terms of their role in public values (Helberger et al., 2018; van Dijck et al., 2018). The EU preliminary principle demands that the MC sector can only be "compliant" in the presence of an oversight body with a transparent system of compliance, an appeal (redress) and a complaints procedure. Any such system would also have to acknowledge and interface with legacy governance structures in the MC sector. An important first step will be to ensure that any advisory or co-regulatory body must include legacy media, community media and citizen representatives to balance the current dominance of computer, digital and telecommunications stakeholders.

The MC sector, especially online intermediaries, should be encouraged to set up an architecture that empowers users. Standardised methodologies and deliberation fora for facilitating ongoing exchange with specific user communities should be put in place. Also, media production cycles, such as designing websites, should involve multiple stakeholders, who should then be required to consider *diversity, non-discrimination, fairness and human rights,* as online game developers were by the ISFE-Council of Europe guidelines (Directorate General of Human Rights and Legal Affairs, 2008). We recommend incentivising and developing educational trajectories, guidelines, training, materials and tools for professional and technical staff (e.g. via online courses or curriculum changes in higher education) among the respective stakeholders to better understand and engage with the EU's fundamental human rights and the principle of trustworthy human-centred AI.

Any system that might emerge should consider the stance some countries have taken in relation to the "traditional media" industries: the Press Councils and Press Ombudsman in Ireland that oversee both print and online only news media (Press Council of Ireland, 2020) and the communications regulation bodies like Ofcom in the UK which oversee telecoms and broadcast media (Ofcom, 2020). Any governance system might also need to work with established worker unions like the National Union of Journalists, both to train and educate journalists, and for whistleblowing and worker rights. In sum, we recommend strengthening workers' rights and public interest values in the media as new AI systems evolve and emerge.

Final reflections

Consideration for the specificities of the MC sector and media examples are largely missing from current EU AI policy documents, as well as in the proposed AIA. This is surprising given high profile scandals involving the misuse of AI in the sector, including the Cambridge Analytica scandal. Further, media experts have been largely absent from key European AI policy initiatives since 2018, and thus we argue that critical social, cultural, democratic and power related issues related to the use of AI in the MC sector are insufficiently considered. Transnational computing, telecommunications and digital platform companies, industry trade bodies and academics have tended to dominate EU policy initiatives in this space. Indeed, our own experience of the AI4People initiative concurs with this, and the expert committee was quite 'thin' in terms of the diversity of stakeholders involved (Raymond & DeNardis, 2015).

Our analysis has proposed that AI systems have a substantial and multi-level impact across the European MC sector. This article identified four levels in which AI applications operate in the MC sector: *automating data capture and processing, automating content generation, automating content mediation* and *automating communication*. We analysed the core uses and risks of AI applications across these levels, identifying where trustworthy AI principles came to the fore, and we also identified the use of emotion recognition AI as high-risk.

It is clear that existing European charters and legislation have not been sufficiently applied to deal with the emergence of a complex socio-technical system of datafication that is built around the technical, economic and political power of a small number of very large commercial transnational corporations operating at both the national and European level. The shift from trustworthy AI to a risk-based but self-administered compliance based system in the AIA does not – in our opinion – auger well for the protection of fundamental rights, freedoms and public values. Neither is it sufficient to protect environmental or social sustainability. The AIA, as currently written, appears to exist in isolation from other policies addressing the digital media sector, and future work will need to address the relationship between AI and other EU digital policies, including the Digital Services Act package.

In concluding our paper, we offered three key areas where more research and regulatory efforts are needed to influence the development, deployment, use and governance of AI systems in the MC sector. These efforts need to involve a range of disciplines and be cognisant of the existing European policy context and sectoral

specifics. The first related to the need to address data power asymmetries by re-thinking informed consent in the context of ML powered AI, and in the deployment and use of predictive analytics by a complex ecosystem of companies. This will also require new systems of transparency and oversight for the use of automated decision making and the dominance of a small number of commercial companies in many of these technologies. The second relates to the need to develop systems that empower citizens and that operate in the public interest, from algorithmic registries to algorithmic auditing, and from skill development to the right to redress. Finally, we consider it necessary to move away from multi stakeholder initiatives towards new systems of cooperative responsibility and accountability to protect citizen and worker rights. While the goal of a principles based and value-driven AI governance framework across the EU is to be commended, too few voices and sectors currently dominate the policy agenda.

References

ADAPT. (2021). *Digital Content Transformation*. Research Centre for AI-Driven Digital Content Technology. https://www.adaptcentre.ie/case-studies/digital-content-transformation/

Balkin, J. M. (2018). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UC Davis Law Review*, *51*, 1149–1210.

Beraldo, D., Milan, S., Vos, J., Agosti, C., Nadalic Sotic, B., Vliegenthart, R., Kruikemeier, S., Otto, L., Vermeer, S., Chu, X., & Votta, F. (2021). Political advertising exposed: Tracking Facebook ads in the 2021 Dutch elections. *Internet Policy Review [Opinion Piece]*. https://policyreview.info/articles/news/political-advertising-exposed-tracking-facebook-ads-2021-dutch-elections/1543

BEUC. (2020). *BEUC's response to the European Commission's white paper on artificial intelligence* (pp. 1–20). BEUC. The European Consumer Organisation. https://www.beuc.eu/publications/beuc-x-2020-049_response_to_the_ecs_white_paper_on_artificial_intelligence.pdf

Binns, R., Zhao, J., Van Kleek, M., & Shadbolt, N. (2018). *Measuring third party tracker power across web and mobile*. https://doi.org/10.48550/ARXIV.1802.02507

Burke, R., Sonboli, N., & Ordonez-Gauger, A. (2018). Balanced neighborhoods for multi-sided fairness in recommendation. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research*, *81*, 1–13. http://proceedings.mlr.press/v81/burke18a/burke18a.pdf

Cassauwers, T. (2019, April 15). Can artificial intelligence help end fake news? *Horizon. The EU Research & Innovation Magazine*. https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news

Chandler, S. (2020, February 7). Reuters uses AI to prototype first ever automated video reports. *Forbes*. https://www.forbes.com/sites/simonchandler/2020/02/07/reuters-uses-ai-to-prototype-first-ever-automated-video-reports/#7eb6a99f7a2a

Charter of fundamental rights of the European Union, 2012/C 326/02 (2000). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, *28*(6), 164–181. https://doi.org/10.1177/0263276411424420

Coalition to Fight Digital Deception. (2021). *Trained for deception: How artificial intelligence fuels online disinformation* (pp. 1–29) [Report]. The Anti-Defamation League, Avaaz, Decode Democracy, Mozilla and New America's Open Technology Institute. https://foundation.mozilla.org/en/campaigns/trained-for-deception-how-artificial-intelligence-fuels-online-disinformation

Content Personalisation Network. (2020). *The Content Personalisation Network project (CPN)*. https://www.projectcpn.eu

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Directorate General of Human Rights and Legal Affairs. (2008). *Human rights guidelines for online games providers. Developed by the Council of Europe in co-operation with the Interactive Software Federation in Europe* (pp. 1–15) [Guidelines]. Council of Europe & IFSE. https://rm.coe.int/16805a39d3

EDPB. (2020). *Guidelines 05/2020 on consent under Regulation 2016/679* (pp. 1–33) [Guidelines]. European Data Protection Board. https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf

Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor* (First Edition). St. Martin's Press.

European Commission. (2018). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe* (Communication COM(2018) 237 final; pp. 1–19). European Commission. https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2018)237&lang=en

European Commission. (2019). *High level expert group on artificial intelligence* [Digital Strategy]. Shaping Europe's Digital Future. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

European Commission. (2020). *White paper on artificial intelligence. A European approach to excellence and trust*(White Paper COM(2020) 65 final; pp. 1–26). https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council, laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act – AI Act) and amending certain Union Legislative Acts* (COM/2021/206 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Commission. (2022). *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)* (COM/2022/496 final; pp. 1–28). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496

European Council. (2017). *European Council meeting (19 October 2017)—Conclusions* (Meeting Conclusions EUCO 14/17; pp. 1–10). https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf

European Parliament & Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119/1 (General Data Protection Regulation)* (Document 32016R0679).

European Parliament. Directorate General for Parliamentary Research Services. (2019a). *Regulating disinformation with artificial intelligence: Effects of disinformation initiatives on freedom of expression and media pluralism*. Publications Office. https://data.europa.eu/doi/10.2861/003689

European Parliament. Directorate General for Parliamentary Research Services. (2019b). *Understanding algorithmic decision-making: Opportunities and challenges*. Publications Office. https://data.europa.eu/doi/10.2861/536131

Fanni, R., Steinkogler, V. E., Zampedri, G., & Pierson, J. (2022). Enhancing human agency through redress in artificial intelligence systems. *AI & Society*. https://doi.org/10.1007/s00146-022-01454-7

Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior*, *116*, 106626. https://doi.org/10.1016/j.chb.2020.106626

Felini, D. (2015). Beyond today's video game rating systems: A critical approach to PEGI and ESRB, and proposed improvements. *Games and Culture*, *10*(1), 106–122. https://doi.org/10.1177/1555412014560192

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, *33*(3), 523–539. https://doi.org/10.1007/s13347-019-00378-3

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

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). https://doi.org/10.1177/2053951719897945

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hagey, K., & Horwitz, J. (2021, September 15). Facebook tried to make its platform a healthier place: It got angrier instead. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

Hasselbalch, G. (2020). Culture by design. *First Monday*. https://doi.org/2013

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism*, *7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

Helberger, N., Eskens, S. J., Drunen, M. Z., Bastian, M. B., & Möller, J. E. (2019). Implications of AI-driven tools in the media for freedom of expression. *Artificial Intelligence – Intelligent Politics: Challenges and Opportunities for Media and Democracy, Nicosia, Cyprus*, 1–36. https://hdl.handle.net/11245.1/64d9c9e7-d15c-4481-97d7-85ebb5179b32

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/10.1080/1369118X.2016.1271900

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, *34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Hintz, A., Dencik, L., & Wahl-Jorgensen, K. (2019). *Digital citizenship in a datafied society*. Polity Press.

HLEG. (2019a). *A definition of AI: Main capabilities and scientific disciplines* (Shaping Europe's Digital Future, pp. 1–7) [Definition developed for the purpose of the deliverables of the High-Level Expert Group on AI]. The European Commission's High-Level Expert Group on Artificial Intelligence. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

HLEG. (2019b). *Ethics guidelines for trustworthy AI* (Shaping Europe's Digital Future, pp. 1–39) [Guidelines]. Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

HLEG. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment* (Shaping Europe's Digital Future) [Assessment list]. Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Information Commissioner's Office. (2019). *Update report into adtech and real time bidding* (pp. 1–25) [Report]. ICO. Information Commissioner's Office. https://ico.org.uk/media/about-the-ico/documents/2615156/adtech-real-time-bidding-report-201906-dl191220.pdf

Irish Council for Civil Liberties. (2021, June 15). ICCL lawsuit takes aim at Google, Facebook, Amazon, Twitter and the entire online advertising industry. *ICCL Press Release*. https://www.iccl.ie/news/press-announcement-rtb-lawsuit/

IVOW. (2020). *An AI and storytelling startup*. IVOW AI. https://www.ivow.ai

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94. https://doi.org/10.1145/2611567

Kennedy, H., Poell, T., & van Dijck, J. (2015). Data and agency. *Big Data & Society*, *2*(2), 205395171562156. https://doi.org/10.1177/2053951715621569

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720915939

Kerr, A., Musiani, F., & Pohle, J. (2019). Communication and internet policy: A critical rights-based history and future. *Internet Policy Review*, *8*(1). https://doi.org/10.14763/2019.1.1395

Lacoma, T. (2020, February 13). League of Legends survey reveals nearly every layer has been harassed. *Screenrant*. https://screenrant.com/league-legends-survey-harassment-toxicity-riot-games-everyone/

Leibowicz, C. (2019, December 11). On AI & media integrity: Insights from the deepfake detection challenge. *Partnership on AI*. https://www.partnershiponai.org/on-ai-media-integrity-insights-from-the-deepfake-detection-challenge/

Lindén, C. G., & Tuulonen, H. (2019). *News Automation. The rewards, risks and realities of 'machine journalism'* (pp. 1–53) [Report]. WAN-IFRA. https://cris.vtt.fi/ws/portalfiles/portal/23705408/WAN_IFRA_News_Automation_FINAL.pdf

Löblich, M., & Venema, N. (2021). Echo chambers: A further dystopia of media generated fragmentation. In G. Balbi, N. Ribeiro, V. Schafer, & C. Schwarzenegger (Eds.), *Digital roots: Historicizing media and communication concepts of the digital age* (pp. 177–192). De Gruyter. https://doi.org/10.1515/9783110740202

MacCarthy, M., & Propp, K. (2021, May 4). Machines learn that Brussels writes the rules: The EU's new AI regulation. *Brookings Techtank*. https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/

Marr, B. (2018, November 12). The amazing ways Google and Grammarly use artificial intelligence to improve your writing. *Forbes*. https://www.forbes.com/sites/bernardmarr/2018/11/12/the-amazing-ways-google-and-grammarly-use-artificial-intelligence-to-improve-our-writing/?sh=3d0841ad3bb0

Meacham, S., Isaac, G., Nauck, D., & Virginas, B. (2019). Towards explainable AI: Design and development for explanation of machine learning predictions for a patient readmittance medical application. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Intelligent computing* (Vol. 997, pp. 939–955). Springer International Publishing. https://doi.org/10.1007/978-3-030-22871-2_67

Media Literacy Ireland. (2020). *About*. Be Media Smart. An Initiative of Media Literacy Ireland. https://www.bemediasmart.ie/about

Mele, C., Russo Spena, T., Kaartemo, V., & Marzullo, M. L. (2021). Smart nudging: How cognitive technologies enable choice architectures for value co-creation. *Journal of Business Research*, *129*, 949–960. https://doi.org/10.1016/j.jbusres.2020.09.004

Milan, S., & Agosti, C. (2019). Personalisation algorithms and elections: Breaking free of the filter bubble. *Internet Policy Review [Opinion Piece]*. https://policyreview.info/articles/news/personalisation-algorithms-and-elections-breaking-free-filter-bubble/1385

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

Moltzau, A. (2020, October 1). Algorithm registries in Amsterdam and Helsinki. *Medium*. https://alexmoltzau.medium.com/algorithm-registries-in-amsterdam-and-helsinki-c1364b70ca6

Napoli, P. M. (2019). *Social media and the public interest: Media regulation in the disinformation age*. Columbia University Press.

Newman, N., Fletcher, R., Schulz, A., Andı, S., Robertson, C. T., & Nielsen, R. K. (2021). *Reuters Institute Digital News Report 2021* (pp. 1–163) [Report]. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021

Nieborg, D. B., & Helmond, A. (2019). The political economy of Facebook's platformization in the mobile ecosystem: Facebook Messenger as a platform instance. *Media, Culture & Society*, *41*(2), 196–218. https://doi.org/10.1177/0163443718818384

Ofcom. (2019). *Use of AI in online content moderation* (pp. 1–82) [Report]. Cambridge Consultants. https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation

Ofcom. (2020). *TV, radio and on-demand*. Ofcom. https://www.ofcom.org.uk/tv-radio-and-on-demand

Papakyriakopoulos, O., Serrano, J. C. M., & Hegelich, S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems. *Online Social Networks and Media*, *15*. https://doi.org/10.1016/j.osnem.2019.100058

Peiser, J. (2019, May 2). The rise of the robot reporter. *The New York Times*. https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html

Pierson, J. (2021). Digital platforms as entangled infrastructures: Addressing public values and trust in messaging apps. *European Journal of Communication*, *36*(4), 349–361. https://doi.org/10.1177/02673231211028374

Pierson, J., & Milan, S. (2017, July 17). Empowerment by design: Configuring the agency of citizens and activists in digital infrastructure. *IAMCR 2017 Conference 'Transforming Culture, Politics & Communication: New Media, New Territories, New Discourses', Cartagena, Colombia*. https://researchportal.vub.be/en/publications/empowerment-by-design-configuring-the-agency-of-citizens-and-acti

Pierson, J., Robinson, C., Boddington, P., Chazerand, P., Kerr, A., Milan, S., Verbeek, F., Kutterer, C., Nerantzi, E., & Crossick, E. (2021). *AI4People — AI in Media and Technology sector: Opportunities, risks, requirements and recommendations* (pp. 212–249) [Report]. Atomium - European Institute for Science, Media and Democracy (EISMD). https://ai4people.eu/wp-content/pdf/AI4People7AIGlobalFrameworks.pdf

Press Council of Ireland. (2020). *Office of the Press Ombudsman*. https://www.presscouncil.ie/

Raymond, M., & DeNardis, L. (2015). Multistakeholderism: Anatomy of an inchoate global institution. *International Theory*, *7*(3), 572–616. https://doi.org/10.1017/S1752971915000081

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879

Scott, M., & Manancourt, V. (2020, September 21). Google and data brokers accused of illegally collecting people's data: Report. *Politico*. https://www.politico.eu/article/google-and-data-brokers-accused-of-illegally-collecting-data-report/amp/

Srnicek, N. (2017). *Platform capitalism*. Polity.

Statt, N. (2020, May 5). Twitter tests a warning message that tells users to rethink offensive replies. *The Verge*. https://www.theverge.com/2020/5/5/21248201/twitter-reply-warning-harmful-language-revise-tweet-moderation

Taddeo, M., McNeish, D., Blanchard, A., & Edgar, E. (2021). Ethical principles for artificial intelligence in national defence. *Philosophy & Technology*, *34*(4), 1707–1729. https://doi.org/10.1007/s13347-021-00482-3

The Guardian. (n.d.). The Cambridge Analytica files. *The Guardian*. https://www.theguardian.com/news/series/cambridge-analytica-files

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Van Zeeland, D. J., Ranaivoson, H. R., Hendrickx, J., Pierson, J., Van den Broeck, W., & Van Der Bank, J. (2019). *Salvaging European media diversity while protecting personal data* (SMIT Policy Brief No. 23). SMIT research group, Vrije Universiteit Brussel. https://cris.vub.be/ws/portalfiles/portal/49332164/

POLICY_BRIEF_23_20180218.pdf

Veale, M. (2020). A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. *European Journal of Risk Regulation*, *11*(1). https://doi.org/10.1017/err.2019.65

Vermeulen, M. (2019). *Online content: To regulate or not to regulate—Is that the question?* (pp. 1–11) [Issue paper]. Association for Progressive Communications. https://www.apc.org/en/pubs/online-content-regulate-or-not-regulate-question

von der Leyen, U. (2019). *A Union that strives for more: My agenda for Europe. Political guidelines for the next Commission (2019-2024)* (pp. 1–22) [Guidelines]. https://ec.europa.eu/info/files/political-guidelines-new-commission_en

Warren, T. (2021, September 7). Microsoft Start is a personalised news feed designed for Windows 11, mobile, and more. *The Verge*. https://www.theverge.com/2021/9/7/22660483/microsoft-start-news-feed-windows-11-features

Whittaker, M. (2021). The steep cost of capture. *Interactions*, *28*(6), 50–55. https://doi.org/10.1145/3488666

Willens, M. (2019, January 3). Forbes is building more AI tools for its reporters. *Digday*. https://digiday.com/media/forbes-built-a-robot-to-pre-write-articles-for-its-contributors/

Winseck, D. (2019). Media concentration in the age of the internet and mobile phones. In M. Prenger & M. Deuze (Eds.), *Making media: Production, practices, and professions* (1st ed., pp. 175–190). Amsterdam University Press. https://doi.org/10.1017/9789048540150.013

Winseck, D. (2020). *Growth and upheaval in the network media economy in Canada, 1984-2019* [Report]. Canadian Media Concentration Research Project (CMCRP), Carleton University. https://doi.org/10.22215/cmcrp/2020.1

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power* (First edition). PublicAffairs.

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies