Responsible Artificial Intelligence

An overview of human rights' challenges of Artificial Intelligence and media literacy perspectives in the context of Bosnia and Herzegovina

www.coe.int/freedomofexpression

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

This research has been done with the support of the Council of Europe and its action “Media and information literacy: for human rights and more democracy”, within the framework of the Council of Europe Action Plan for Bosnia and Herzegovina 2018-2021. Action Plan level funding is provided by Luxembourg and Norway.

Cover design and layout: Information Society Department, Council of Europe

Photos: ©Shutterstock

This publication has not been copy-edited by the SPDP Editorial Unit to correct typographical and grammatical errors.

Responsible Artificial Intelligence

An overview of human rights' challenges of Artificial Intelligence and media literacy perspectives in the context of Bosnia and Herzegovina

February 2022

Authors:

Bojana Kostić
is a long-time advocate for media freedom with an extensive background in community-led and inclusive research. In her research, she explores issues that lie at the intersection of human rights and technology, focusing on freedom of expression online, online abuse, safety and silences.

Caroline Sinders
is a critical designer and artist. For the past few years, she has been examining the intersections of artificial intelligence, intersectional justice, systems design, harm, and politics in digital conversational spaces and technology platforms.

Council of Europe

Table of Content

Abbreviations

AAAI	The Association for the Advancement of Artificial Intelligence
AARP	American Association of Retired Persons
ADS	Automated Decision System
AI	Artificial Intelligence
AVM	Audio and Visual Media
BiH	Bosnia and Herzegovina
CAHAI	Council of Europe Ad Hoc Committee on Artificial Intelligence
CoE	Council of Europe
CoM	Committee of Ministers
CRA	Communication Regulatory Agency
Decl	Declaration
DMA	Digital Markets Act
DSA	Digital Services Act
EC	European Commission
ECI	European Citizens' Initiative
EDRi	European Digital Rights
EFF	Electronic Frontier Foundation
EU	European Union
FSLN	Sandinista National Liberation Front
GDPR	General Data Protection Regulation
GIFCT	Global Internet Forum to Counter Terrorism
HITL	human in the loop
IBM	International Business Machines Corporation
IEEE	Institute of Electrical and Electronics Engineers
KRIK	Crime and Corruption Reporting Network
MIL	Media and Information Literacy
OSCE	Organization for Security and Co-operation in Europe
PACE	Parliamentary Assembly of the Council of Europe
Rec	Recommendation
RFoM	OSCE Representative on Freedom of the Media
SAIFE	Spotlight on Artificial Intelligence and Freedom of Expression
UNHRC	United Nations Human Rights Council

UNISYS	United Information Systems
UvA	University of Amsterdam
WB	Western Balkan

I. Introduction

In the public and private sectors, both online and offline, artificial intelligence (AI) and algorithms make important decisions for people's lives. For this reason, AI has generated global policy attention aimed at addressing the human rights challenges posed by artificial intelligence. This study on Responsible Artificial Intelligence explores AI's multipronged impact on human rights, specifically investigating the intersection of AI and human rights through the broad lens of freedom of expression and media freedom.

Large technology companies (social media platforms, streaming and sharing platforms, and most notably Facebook, Google, Twitter, Tik-Tok, Netflix, Amazon, Apple, Microsoft, IBM, Samsung, etc.) are gaining unprecedented economic and political power. The increased development and utilisation of AI and algorithms amplifies their societal impact and directly affect individual human rights, democracy, and the rule of law (CM Decl(13/02/2019)1, para. 9). Unpacking, explaining, and situating the effects of AI is necessary to create new forms of responsibility frameworks and new media and information literacy (MIL) interventions. Against this backdrop, the objectives of this study are twofold: i. to introduce stakeholders and the expert community to the practical implications of AI's societal impact; ii. to propose resources for MIL in the context of the particular AI's impact.

This study consists of two conceptual components, which explore:

i) The power of the automated systems within the daily lives of people in the digital environment, focusing on the following concepts:

- impact of AI on people's decision makings
- digital power
- user's agency
- data repurposing
- deepfakes
- algorithmic harms.[1]

ii) The impact of AI on freedom of expression and media freedom in regard to popular and global platforms, software, and applications.

By combining in-depth academic literature review with extensive MIL resources compilation, this study provides critical insight into AI's current landscape and the challenges it poses on human rights. It should be noted that this study does not cover all relevant aspects of AI and human rights as many AI related topics fall outside its scope – such as sector specific use of AI,[2] data protection concerns outside of AI and freedom of expression, as well as oversight and human rights assessment standards. Some of these aspects are mentioned as illustrative examples or supporting arguments, but the focus remains on articulating and mapping out the current expert and regulatory discussions through the lens of AI and freedom of expression.

Grounded in understanding the role of MIL and active citizenship, this study intentionally focuses on large technology companies like Facebook because of the public and academic attention they have recently received (Newton 2021). Additionally, Facebook is the most used

[1] Note: there is no universally accepted or authoritative definition of these terms. They are context dependent, open-ended and fluid.

[2] Such as good governance and rule of law, financial system, etc.

social media platform in Bosnia and Herzegovina (BiH). In a country where 96% of the population has internet access, over three-fourths (78%, CoE 2021, p. 5) have social media accounts, 71% (ibid. p. 6) of which are Facebook users. While not as popular as Facebook, video-sharing platforms are used by 42% of BiH's internet users (ibid., p. 5). Compared to adult internet users, young people tend to spend more time online – on average four hours (ibid., p. 5) – and are more likely to passively consume content that is algorithmically recommended (Hodžić 2019, p. 32, 34). However, the younger population is not the only community exposed to the risks of automated content. Internet users, in general, are more likely to read, than generate and share new content and comments (CoE 2021, p. 6). Consequently, online public deliberation is relatively low (46%, ibid.), even though social media are mostly used as forums for public discussion (ibid.). As a highly internet-connected country, social media platforms (52%, ibid.) and online news periodicals (45%, ibid.) are important sources of public information and deliberation, but as seen from the findings of the Study on Media Habits and Attitudes of Adults in Bosnia and Herzegovina (CoE 2021), most users are passively consuming this content.

It is important to note that there is little to no information on the country's state and non-state actors' relations with these large technology companies (Kostić (forthcoming)). The increasing number of problems arising on social media such as hate speech, polarisation, shrinking online civic spaces, political and inter-ethnic hate remain unaddressed (Turčilo et al. (forthcoming); Cvjetićanin 2019, p. 7, 22, 40, 63; Sokol 2020, p. 20). In this post-conflict country, these unresolved issues resonate differently and colour public discourse and the prospects for peace and societal cohesion. For this reason, among many others, MIL interventions that offer citizens ways to critically engage with AI and algorithms and understand their relevance and impact on freedom of expression online, are vital in fostering the country's democratic culture and societal prosperity.

To understand the challenges AI and algorithmic systems pose on human rights, most notably freedom of expression, and media freedom, this study provides a critical assessment of the current digital landscape and the responsibility frameworks of making socio-technological processes transparent and accountable. Chapter II defines and explains the basic AI concepts explored in this study. Coupled with illustrative examples, the following Chapter III offers an analysis of key AI concepts. Chapter IV provides an overview of the challenges AI and algorithms pose to freedom of expression and the emerging tensions between media freedom and large technology companies. A brief regulatory map is presented in Chapter V with emphasis on MIL interventions. This study concludes by reiterating the relevance of AI and algorithmic impact on our future, and in addition, it presents a set of useful MIL resources (Annex A) and a glossary of terms (Annex B), which should assist in developing local MIL and AI interventions.

II. Basic AI vocabulary

AI and algorithmic decision-making have emerged as umbrella terms, thus there is no consensus regarding their definitions (Zuiderveen Borgesius 2018, p. 11), nor uniform classification of the concepts emerging in this field. AI, algorithms, and machine learning are used interchangeably across the industry, academia, journalism, and civil society. Technically, each of these terms means different things; but colloquially, they are used synonymously. The common denominator that envelopes these technological expressions under a common

terminological umbrella is their task-oriented functionality, large-scale datasets,[3] their technical structures, and their components. To provide a more cohesive comprehension of the language used in this study, this Chapter breaks down basic and key AI concepts, as following:

Algorithms. An algorithm can be described as "an abstract, formalised description of a computational procedure and [...] as a rough rule of thumb, one could think of an algorithm as a computer program" (Zuiderveen Borgesius 2018, p. 11) or as "a series of steps (or set of rules) for solving or performing a task" (Onuoha & Nucera 2018, p. 8). The Council of Europe glossary on AI describes algorithms as a "finite suite of formal rules (logical operations, instructions) allowing to obtain a result from input elements. This suite can be the object of an automated execution process and rely on models designed through machine learning" (CoE AI Glossary). While some algorithms carry out tasks automatically, like in the case of spam or hashtag filters that remove online hate speech and extremist language (GIFCT - Share Industry Database), others assist human moderators in making decisions regarding content removal. For example, algorithms will flag problematic content, but human moderators will make the final decision about keeping or omitting the flagged content (Kostić 2021, p. 23).

Artificial intelligence (AI) and machine learning. Artificial intelligence, loosely speaking, is "the science of making machines smart" (Zuiderveen Borgesius 2018, p. 12). The Council of Europe glossary defines AI as "a set of sciences, theories and techniques whose purpose is to reproduce, by a machine, the cognitive abilities of a human being to be able to entrust a machine with complex tasks previously delegated to a human" (CoE AI Glossary). International organisations, such as the Organization for Security and Co-operation in Europe (OSCE), define and explain AI as "based on algorithms, which are sets of human-designed instructions with encoded procedures for transforming input data into a desired output, based on specific calculations" (OSCE SAIFE 2020, p. 27). The over 100-year-old technology company IBM defines AI as "artificial intelligence [that] leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind" (IBM Cloud Education 2020).

In sum, algorithms are the building blocks and thus an integral part of AI. Some algorithms have the possibility of so-called computer or machine learning. This type of (semi) autonomous learning can be described as a developmental AI technique designed to improve the quality of automated decision making, by recognising patterns and "regularities" to carry out certain tasks independent of human intervention – in other words, the possibility of "learning without explicit programming" (Privacy International & Article 19 2018, p. 7). This process is also referred to as machine learning or "a branch of artificial intelligence in which a computer generates rules and predictions based on raw data that has been fed into it" (Onuoha & Nucera 2018, p. 8).

However, the term *learning* is a misnomer as "the computer is able to find similarities and differences in the data through the repetitious tuning of its parameters" (Leslie et al. 2021, p. 8). Machine learning relies on classifier structures that enable machines to make a set of assumptions and thus making them prone to error. While machine learning algorithms are able to perform repetitive computational tasks, they are not able to correct the errors of their assumptions without human intervention. Their repetitive nature, therefore, mirrors the ability to regurgitate rather than comprehend information (Leslie et al. 2021, p. 8-10).

[3] "Large-scale datasets" refers to datasets collected and created by technology companies to be used by AI systems. To train and generate AI models, AI systems require hundreds of thousands of data points within datasets.

This autonomous and misleading perception of human-like intelligence has generated mass public and expert attention, which has further mystified these computational processes. Regardless of this veil of mysticism, AI is not a neutral but an active social agent (EC 2020, p. 24). AI is written, created, and coded by people. The computational code and datasets that fuel algorithms, which in turn are the building blocks of AI, reflect and refract human values, biases, wants, and desires. The human engineers who create these systems are actors situated within networks of wealthy and powerful corporations, and social media giants that exploit these systems at the cost of the country's democracy for their own economic gain. Therefore, human values, biases, wants, and desires are enlaced and faintly threaded in the very fabric that is AI – on an individual (human engineer) level and systemic organisational and institutional level (Mazzoli & Tambini 2020, p. 30).

The process of developing AI reflects thousands of individual biases situated within the initiatives of the companies that employ them. These multidimensional entanglements (on a micro and macro level) must be recognised. This study uses terms such as AI, and algorithms separately as it is crucial to ensure etymological, epistemological and practical separation.

AI through illustrations

Algorithms are part of every computer system, they are the basic elements of almost every digital artefact – from recommendation systems on social media, audio and video media (AVM) services, to a myriad of health, fitness, and mental support apps. For example, AI sometimes requires more sophisticated and complex algorithmic methods currently used in technologies like AI chatbots – "conversational agent that dialogues with its user" (CoE AI Glossary) such as Microsoft's Tay that tweets like a teen (The Verge 2016). The language processing and complex translation operations – Google Perspective, which can detect toxic comments, also require more sophisticated algorithms. This is also the case for a number of surveillance technologies (see: Reclaim your face, ECI 2021) as well as biotechnologies that are currently deployed to control and counter refugee influx on the borders of the European Union (Molnar 2020).

Automated decision-making systems. In 2021 the Alan Turing Institute published a paper on AI, human rights, democracy, and rule of law that described automated decision systems (ADS) as technological processes that augment and replace human decision-making processes with computer processors to answer different questions such as "discrete classification" (e.g., female-male-non-binary) or assess score like in cases of creditworthiness and risks crime occurrences (Leslie et al. 2021, p. 36). These systems mainly rely on the use of "trained" datasets that were previously programmed to look for similar data correlation and points (Ackerman 2021). Often, more complex ADS do not require any human intervention, even though they are able to automate decisions and choices that affect a person's life. Due to the lack of meaningful transparency, public scrutiny, and accountability of the actors that create these technologies, ADS's underlying issues often fall into one of two dimensions: internal or external.

In relation to *internal* issues, companies often build technological infrastructure and deploy ADS without taking into consideration their limitations or societal and political implications. As the public interacts with these unaudited systems, they struggle to understand the technological nuances they encounter, which makes it difficult to supply meaningful intervention. This is witnessed in AI systems designed to detect fraudulent behaviour. Instead of facilitating the process of receiving government benefits, these systems "mistakenly" prevented people who were in need of them the most (Gilman 2020). The root of these internal

issues lies in the datasets used to "train" these systems. The datasets that inform the decisions ADSs make are riddled with inherent "anomalies" that lead to inadequate or skewed results. The problem comes down to: i. the lack of data; ii. the lack of diversified data (e.g., only providing gender data in binary terms); and iii. inaccurate and low-quality data (Kostić 2021, p. 14; for extensive discussion, see: EC 2020, p. 27-30). Ultimately, these internal issues explain how facial recognition systems of large technology companies have a harder time accurately recognising race and gender (Leufer 2021).

It is crucial to understand that all datasets are bias and the process of developing these systems is inherently driven by the values and goals of the people who build and use them. To ensure that the design and process of developing these technologies account for their potential harms and risks, and to recentre the needs of human rights and individual autonomy, we must ask ourselves: *who are these systems failing?*

Therefore, to address *the external* issues of automated decision-making, requires clear, affordable, effective, and accessible redress mechanisms that enable individuals to report "problems" (CoE 2019, p. 13, CoE MSI 2018, para. 4.4. and 4.5). To that end, the concept of *human in the loop*[4] (HITL) (EC 2020, p. 32, see also: CoE 2019, p. 9-10) refers to the need for human intervention within these automated processes to carry out oversight and corrective functions to ensure that individuals' human rights are indeed respected and protected. However, humans are not homogenous and different people also have different needs and experiences.

Hence, another important concept, *society in the loop*[5] (Rahwan 2017, p. 3; CoE 2019, p. 11), proposes to highlight the values and interests of different societal groups, ethical and human right principles, and participatory opportunities in the design, development (internal dimension), implementation, and accessibility of the redress mechanisms (external dimension) of the automated systems for decision making (CoE Conference Conclusion 2021, para. 8). When properly developed and designed, ADSs can have a positive impact. For example, the inclusion of recommending systems that intentionally elevate content that promotes peace and diversity.

Responsible AI. Addressing harms and their impact on individuals and groups essentially means exploring "values embedded in algorithms that need to be questioned, critiqued and challenged. Indeed, it is not the algorithms themselves but the decision-making processes around algorithms that must be scrutinised in terms of how they affect human rights" (Wagner et al. 2018, p. 8). An increasing number of initiatives and actors are involved in the process of understanding and countering these harms, which has given rise to an emerging field: responsible AI. This field offers guidelines for the design, development, and deployment of AI that align with ethical and human rights standards (e.g., the Association for the Advancement of Artificial Intelligence (AAAI)). This framework encompasses a set of principles and requirements that centre human responsibility over any decision affecting individual rights and freedoms within the proposed accountability and liability framework. However, this is easier said than done. The design and implementation of Responsible AI entails "dealing with

[4] Note: "human in the loop (HITL) refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impacts) and the ability to decide when and how to use the system in any particular situation." (EC 2020, p.32).

[5] Society in the loop is a nod to a common AI term and system set up of 'human in the loop' where a human is integrated and is a part of the AI system, as a form of AI training, data integrity analysis, or helping refine the AI model (Appen 2019).

imperfection and realising that tensions and dilemmas may occur when "doing the right thing does not have an obvious and widely agreed upon answer" (Rakova et al. 2021, p. 3).

Essentially, integrating notions of responsibility into AI requires the involvement of different stakeholders and their agencies, not just tech-led solutions. In the context of the AI regulatory landscape, we are witnessing multi-disciplinary and multi-stakeholder cooperation and intervention, which is briefly discussed under Chapter V. Regulatory and policy landscape: a brief overview. To understand the need for this kind of cooperation and intervention, the following Chapter unfolds key concepts as they intersect with human rights and individual agency.

III. AI and daily lives: unpacking key concepts

AI and algorithms exist in nearly every facet of a person's life – they affect how we search for information online, the kinds of social media content we view, the media and entertainment content we consume, and the marketplaces we utilise. Threaded in the fabric of our reality, AI is one of the key drivers of change and "a determining factor for the future of humanity as it will substantially transform individual lives and impact on human communities" (PACE 2020, para. 1). The following section unpacks key AI concepts explored in this study, in order to understand how their ubiquitous nature and our increasing co-dependence on these systems affect and shape our individual autonomy.

3.1. Digital power

As they compete for our free labour in the form of user data – interaction, engagement, time, and behaviour patterns (also referred to as "attention economy" or "like" economy) (Gerlitz et al. 2013) – the systems these companies craft and mould make important choices regarding what kind of information is visible and to whom. As they structure the visibility of our digital reality, algorithms and AI become "information gatekeepers" and transform the large technology companies that emerged as "the custodians of the internet" (Gillespie 2018). As they prioritise profit over human rights and individual agency (PACE 2020, para. 6; Maréchal & Biddle 2020), these companies profit from the exploitation and extraction of personal data and behaviour patterns. They appropriate "human life through data work to order economic and social life as a whole", thus manifesting themselves as agents of digital power (Couldry & Mejias 2019, p. 33).

In this study, digital power refers to the concentration of data, information, and influence instilled in a small number of tech-based companies, coupled with the increasing dependence of individuals, societies, and institutions on the provision of their services and the lack of effective democratic oversight (PACE 2020, para. 14.6; Yeung 2019, p. 38). The digital power they hold enables companies to gain not only economic power, but also political power as "social media platforms are embedded in complex governance structures and accountability relationships with a range of different stakeholders: not only governments but also proactive users, civil society actors, and commercial partners may motivate them to intervene in content flows" (Leerssen 2020, p. 10).

In other words, the digital power of large technology companies is not simply a matter of attention economy, but also, a matter of "attention politics" (ibid., p. 11). This economic and

political power is unevenly distributed and further marginalises the digital needs of countries like Bosnia and Herzegovina. Due to their small population, lack of information-technological infrastructure, and/or purchasing power, Bosnia and Herzegovina has not been able to participate as an active actor in the internet globalisation discussion. Thus, despite the country's widespread use and reliance on social media services, the companies that provide these technologies do not have representatives in this country (or in the Western Balkans). Consequently, when individuals, journalists, and other actors face AI and algorithmic related problems they are unable to reach these companies and "solve" the problem (Kostić (forthcoming)).

Digital power and discrimination

The following examples highlight the concentration of digital power over online communication and interactions to illustrate their discriminatory treatment of:

- certain content over other (inflammatory speech is more prominent as it gains more attention and user's engagement, The Markup 2021);
- certain forms of speech over others (under the newsworthiness exception, a standard devised by Facebook to allow political actors (and their speech) to stay online against their internal terms of services (Ohlheiser 2019; Klonick 2017, p. 1665);
- certain languages (there is growing evidence that removal of hate-speech is more "efficient" in assessing the English language, Mozilla 2021, p. 19);
- certain societies and countries over others (the onus of decision and regulatory agenda is set by the United States and EU).

The previous examples and the brief discussion on the digital power of AI reflect – through a so-called "network effect" (European Digital Rights Initiative 2020) – how AI is shifting the power from state representatives and citizens to private and unelected actors (Kalluri 2020, p. 169). In Bosnia and Herzegovina, this shift in power needs to be observed through a complex media landscape lens (EC 2020, p. 5, 17) especially given people's general passivity when consuming digital content. To re-shift this power dynamic and unlock the unlimited potential of public participation in the digital sphere, responsible AI frameworks, including MIL interventions, must secure and ensure individual autonomy and freedom of expression.

3.2. Impact of AI on people's decision making

As noted above, the fabric of our digital society is dependent on AI and algorithms. These systems process our personal data and online behaviour patterns to make assessments and generate outcomes. From estimating a person's weight loss journey to finding less expensive products to complex facial recognition software, AI and data-driven processes are deeply ingrained in the decisions and choices we make, thus affecting our overall individual autonomy – the level of agency we have in making our own decisions. Although the "space" to practice individual autonomy has always intersected with a range of political, economic, and societal factors, in today's environment this "space" is saturated with "automated agents that have the power to shape the contexts in which human agents make choices" (Couldry & Mejias 2019, p. 183). Simply put, when AI decisions impact an individual's decision-making process, that individual's autonomy becomes "outsourced" to large technology companies (ibid., p. 214).

Large technology companies and freedom of expression

In relation to freedom of expression, large technology companies substantially affect what is said and not said, who is and is not seen, and what information one is and one is not exposed to. As witnessed in the following instances, these companies' gatekeeping power curtails freedom of expression:

- free and fair election process (e.g., disruption and interference of the electoral processes (for example, see the Cambridge Analytica case, The Guardian 2019);
- peace prospects (e.g., increased societal polarisation - for example, the case study on Myanmar, UN HRC 2018);
- personal autonomy and decision making (e.g., advertising and opinion nudging through massive profiling, The Mark Up 2021);
- individual behaviour (e.g., dark patterns are deliberately pushing for certain behaviour of the end-users, Sinders 2021);
- societal cohesion (e.g., deep fakes, discrimination against people of colour or indigenous groups, see Noble 2018; PACE 2020, p. 21-23).

Among various risks stemming from these systems, profiling is notoriously problematic. Through profiling, AI technologies regularly infer personal details and behavioural patterns from seemingly uninteresting data to produce digital profiles that may or may not be accurate. This process referred to as data-driven "persuasion profiling" (Couldry & Mejias 2019, p. 140) enables companies to map out a person's cognitive preferences and behaviours to inform the future decisions they make. Therefore, consequential decisions (Privacy International 2017) are shaped by the profiles these systems fabricate for "targeted, personalised and often unnoticed influence on individuals and social groups, which different political actors may be tempted to use to their own benefit" (PACE 2020, para. 4). Although profiling is a relatively novel concept in European data protection law, it is now explicitly defined under Article 4(4) of the EU General Data Protection Regulation (GDPR) and refers to the automated processing of personal data to derive, infer, predict or evaluate information about an individual (or group) to analyse or predict an individual's (or groups) identity, attributes, interests, and behaviours (GDPR, Art. 4(4) and rec. 71).

Inaccurate profiling and downstream risks

In 2017 Forbes contributor Kalev Leetaru, in his 30s, requested information about himself from Oracle, a large data broker, after accidentally receiving a membership from the American Association of Retired Persons (AARP), which is an organisation for United States individuals who are retired or aged 50 and above. Oracle sent Leetaru documents and data about him, including a list of categories they attributed to him. Leetaru found that out of a total of 108 categories Oracle had associated him with, 85 of them (about 78%) were woefully inaccurate. This kind of data Oracle collects is sold to many other companies and if this large amount of data that is gathered and collected is incorrect, then the repurposed data can result in many harmful downstream effects for individuals (Leetaru 2018).

3.3. Data repurposing

Large technology companies represent a concentration of digital power because they operate under a business model motivated by "bulk and mass data collection, analysis and surveillance" (Ranking Digital Rights 2020). The effectiveness of this model is witnessed in these companies' advertising revenues, which are the biggest in the world. Through this model and

related business practises, large technology companies have "locked in" this business model logic and centralised themselves in the lives of individuals, organisations, the media, other companies, and smaller businesses. As individuals become increasingly dependent on the services large technology companies provide, data production becomes decentralised and data collection becomes recentralised (Helmond 2015, p. 12).

In practice, this means that the digital traces individuals leave online (e.g., shares, likes, visits and time spent on websites and social media platforms), which seemingly only exist within digital spaces, are collected, stored, and analysed by these large technology companies. For example, when an individual visits a site for the first time, a file called a cookie is created in their browser's directory to form a unique link between the individual and the site visited. All the actions performed on this site, like adding a product to a cart, are stored within these small files (cookies). While designed to navigate digital spaces more efficiently – such as remembering the shopping cart from a previous visit – keeping these actions on file enables cookies to become digital blueprints to our online traces (CoE AI Glossary). These digital blueprints enable companies to track, trace, and map out our online behaviour. To "combine data about various selves", these digital blueprints are often paired with offline data (Couldry & Mejias 2019, p. 21).

Given their lack of transparency, large technology companies are in a position to repurpose, reuse, and resell our personal data, ad nauseam.[6] These companies utilise their digital power to map out our digital traces by collecting individual data across many different digital spaces, entities, platforms, and processes such as ad tracking (advertisement tracking software that gathers personal insights regarding users' behaviours and preferences to develop targeted ads (Maréchal & Biddle 2020, p. 25-26) or cookies. Although initially collected and processed for a specific purpose (i.e., cookies capturing data to navigate sites more efficiently), this data is readily recycled and used for different purposes. For example, Cambridge Analytica and its parent company, SCL Group, acquired Facebook user data and repurposed it to target individual groups and nudge their voting behaviours, thus having a detrimental effect on elections across 30 different countries (Ghoshal 2018). The ability to repurpose, reuse, and resell our digital traces has given rise to several problematic processes such as data brokers – companies that collect, re-pack, and sell personal data related to different aspects of human life (Couldry & Mejias 2019, p. 52). In December 2013, during a US Congressional hearing, the public caught a glimpse of the data broker world, and the risks it poses. The testimony revealed that data brokers were collecting and selling data from rape and domestic abuse victims (ibid., see also: Dixon 2013).

Discriminatory advertising practices

The investigative journalism outlet, the MarkUp, found that Facebook is allowing companies, like Exxon Mobil, to target different kinds of political groups on Facebook with different kinds of ads. Within this example arise two specific problems, one in which Facebook uses data profiling measures to determine users' political leanings based on the kind of content users interact with and share; and the second in allowing companies to target specific kinds of groups from the particular interfaces that Facebook has identified (Merrill 2021). This example shows the capabilities that are within platforms to target users, and that

[6] Council of Europe Recommendation of the Committee of Ministers to member States on Big Data for culture, literacy and democracy (CM/Rec(2017)8) defines repurposing as "finding a new use for a given object and redeploying it by assigning an alternative use and value to it, or a different format and context, which, in the digital world, implies the creation of metadata and data". (p. 7).

those targets are often created by the platforms themselves, with the users having no knowledge of how they are labelled or taxonomised by these platforms.

3.4. User's control and agency

Data extraction and the lack of agency that users have in controlling this extraction process are prevalent across nearly every industry and country. Within the design and technology industry, this form of agency or "informed consent" – a statement or clear affirmative actions that signify the user's permission to process their personal data (Article 4(11) GDPR) – refers to the kinds of decisions individuals can or cannot make in regard to their personal data and history of digital behaviour. Often technology and products will give the illusion of control or agency through the small decisions an individual is able to make such as how their profile is presented on Facebook or what kinds of content they watch on Netflix. However, in reality, these products offer little meaningful agency. The content a person views from Netflix is selected among an algorithmically curated list of viewing preferences. Similarly, on Facebook a person cannot control which companies access their data or what type of content is made visible in their timeline, nor can they determine what products are advertised to them on Amazon or other shopping platforms.

Target's Targeted Surveillance

The algorithmic content decision-making extends across e-commerce platforms: with Amazon price gouging during the pandemic on necessary items like hand sanitiser (Harrison 2020) to how or why content is shown within a user's algorithmic timeline on Facebook (where Facebook's algorithm prefers to share misinformation and disinformation, Merrill & Oremus 2021). Another interesting example is from 2012 when it was discovered that Target was creating massive databases and data tracking on their consumers to send them coupons and information without their consent. At first glance this does not sound like a malicious or harmful example, however Target was collecting highly specialised information without customers' awareness and consent, through which it was able to discover whether customers were pregnant. The New York Times found that based on the products users bought, Target was able to assign to customers a pregnancy prediction score. From this data, the company was able to predict specific stages of pregnancy and would send out personalised coupons for each stage of pregnancy. With this additional context, it is easier to see how customers felt their personal data and information had been violated and misused (Duhigg 2012).

3.5. Algorithmic harms

The digital power large technologies harness and the lack of individual autonomy do not affect all individuals equally. "The technology-based amplification of bias and prejudice, as well as statistical flaws and errors" (PACE 2020, para. 25) can propel and amplify societal inequalities and discriminatory practices, further marginalising disenfranchised communities and post-conflict societies (Keller 2021). In the context of online freedom of expression, scholars have produced a rich vein of literature that explores and unveils various forms of algorithmic harms such as intersectional discrimination (Noble 2018), amplification censorship (Cobbe 2020, p. 9), informational gaps (Čaušević & Sengupta 2020), a large-scale information manipulation (Nikolic & Jeremic 2020), and insidious private-public partnerships (Feldstein 2020).

However, outside the realm of networked communication, algorithmic harms also reinforce and perpetuate discriminatory practices. For example, a facial recognition software that is not able to recognise different skin colours (see: Section II. Basic AI vocabulary, also see project: Our data bodies). In some instances, the inability to recognise a person's ethnicity and/or race can prove inconvenient such as in cases where buildings rely on this technology to permit residents to enter (The Guardian 2019). In other instances, the harms of relying on systems that are unable to adequately identify a person's characteristics are more obvious. This includes systems of law enforcement that rely on this technology to assist in criminal investigations (Al-Kawaz et al. 2018). While in other cases, – the societal inequalities and discriminatory practices – these systems are ontological such as sociotechnical design grounded in the belief that the many facets of our identity (race, ethnicity, gender, sexuality, etc.) are static and quantifiable variables (see example below).

Automating Gender from Iris

In 2007 the Institute of Electrical and Electronics Engineers (IEEE) published a study funded by UNISYS Corp and several US governmental agencies, including the Central Intelligence Agency (CIA) that proposed machine learning models can predict a person's gender from the texture of their iris (Thomas et al. 2007). These models were trained using images of participants' eyes. The sample image provided in the article reveals that the images included not only the participants' iris but their entire eye (including their upper and lower lashes) (ibid., p. 3). Conducted by a group of male computer scientists and engineering scholars from the University of Notre Dame, this study reported "gender classification models that can reach accuracies close to 80%." (ibid., p. 5). The study, in addition to several others (Tapia et al. 2016; Fairhurst et al. 2015; Tapia et al. 2014, Bansal et al. 2012; Lagree & Bowyer 2011), suggested that machine learning algorithms can predict a person's gender. This was not only a step backwards in understanding that gender as a social construct cannot be quantified, but it also operated under the notion that AI systems are neutral actors.

Ten years later the IEEE published another study conducted (again) by a group of male scholars, one of which is an author of the previously cited study while the other two are from the same department and institution (Kuehlkamp et al. 2017). This study of 2017 stood apart as it acknowledged how previous gender predicting algorithms failed to account for effects such as cosmetics. Using the same dataset as the study conducted in 2007, these scholars retrained the algorithms to account for the effects of makeup and found that the presence of makeup in the images resulted "in higher estimated gender-from-iris accuracy". Effectively, this study revealed that these algorithms were not able to predict gender but instead identify the presence of makeup.

IV. Overview: AI and freedom of expression

A functioning democracy requires open, vibrant, and unhindered public debate and free circulation of information. That is why freedom of expression and media freedom are the lifeblood of democracy. However, the increased dominance and impact of large technology companies, fuelled by algorithms and AI, transform the way we express our views, receive information, and engage in public debate. This radical transformation "captures the power and scope these private platforms wield through their moderation systems and lend gravitas to their

role in democratic culture" (Klonick 2017, p. 1663). For this reason, it is essential to understand these transformative processes and their dynamics. This Chapter looks at the salient aspects of these dynamics by first focusing on the relationship between AI, algorithms, and freedom of expression, and then exploring the effects these systems have on media freedom.

4.1. Content moderation

AI and algorithmic content moderation – an umbrella term for processes of content monitoring, assessment, selection and distribution (Bukovska 2020, p. 32-34) based on profiling and surveillance practises – pose high risks to an individual's ability to reach informed opinions (Leslie et al. 2021, p. 14). To be able to understand the risks, scale, and complexity of this process, the following paragraphs illustrate (in a simplified version) a content moderation cycle.

On social media platforms, every piece of content goes through three different levels of moderation, especially user-generated content, meaning digital media produced and disseminated by individual users (Jenkins 2006; OSCE SAIFE 2020, p. 3).

Level 1: Uploading content

Before content is online, a set of "upload" automatic filters assess the material to determine whether or not it falls under the list of prohibited content for publishing online as defined by the platform's Terms of Services and internal documents (e.g., to prevent child pornography or copyright infringing content).

Level 2: The custodians of the internet

Content that passes the first level of moderation is then evaluated by AI and algorithms. These automated processes are designed to rank, optimise, and recommend content based on a number of criteria and data-points (e.g., the category of content in question, personal profiles and audience preferences). In this stage, automatic processes make decisions about how visible this content is, thus operating as "exposure" architects. Given that there is almost no information about these AI ranking and recommendation systems, they are impenetrable and often referred to as a black box (Pasquale 2015).

Level 3: To be or not to be

Once content is published online, it lives under a set of processes and practises that are "governed" by reporting mechanisms. These reporting mechanisms are established under the platform's Terms of Service, and enable individual users to report, flag, and block categories of content pre-defined by the platform's internal documents. If the content is flagged, a reporting mechanism is triggered, and a combination of AI and human assessments makes decisions about the report. Consequently, reported content (including personal accounts) can be blocked, taken down or otherwise sanctioned.

(In)efficient reporting mechanisms and the Western Balkan experience

Despite these three levels of content moderation system, hateful and harmful content is still widely prevalent in the digital sphere. For example, a study by the Balkan Insight Reporting Network found that almost half of the reported content produced from Western Balkans remains online (Jeremić et al. 2021). This study also found that problematic content flagged by women is less likely to be removed (62% as opposed to 38% of men, ibid.). These reporting mechanisms, which are often referred to as redress mechanisms, must be critiqued for their inefficiency, lack of transparency, and arbitrariness. Our co-dependent relationship

with these technologies needs to be re-evaluated. In this content moderation process, human intervention is not always available, therefore a problem caused by AI and algorithms at one level becomes automated at another level (Bukovska 2020, p. 3).

4.2. Content curation

At the second level of content moderation, the digital power of large technology companies becomes clear. Their ability to control the circulation of global digital content is often referred to as content curation (Gorwa et al. 2020, p. 3; Bukovska 2020, p. 19). For example, a given Facebook newsfeed is produced from processing an individual's digital behavioural patterns (time spent on a similar comment, previous engagement, personal and group profiles that serve as a proxy) to optimise and personalise each piece of content (Constine 2016; Bernstein et al. 2020, p. 47) – which is why there are no two identical Facebook accounts or NewsFeeds.[7]

AI and algorithmic recommender systems make automated and personalised decisions about what the end-user sees on their profile. Given the volume of digital content and the velocity through which it is uploaded, content curation through the use of AI and algorithms, seems indispensable. From a functional perspective, these systems ensure that people are able to navigate these dense informational forests. For example, without knowing an individual's browsing history it would be very hard for an individual to perform a simple internet search (Bodó et al. 2019, p. 207). Therefore, automated content curation systems are not only part of the problem but also invaluable members of the digital assembly line (Burri et al. 2020, p. 42).

Content curation and recommendation processes are not only driven by AI and algorithms, but also by business models and profit logic. The large technology companies that deploy these systems are financially sustained through ensuring individuals spend more time on their platforms to maximise an individual's exposure to advertisements. In doing so, there is an inherent risk that recommendations systems predominantly expose individuals to content that is most likely to attract their interest – also known as internet filter-bubble phenomenon (Bodó et al. 2019, p. 209; Kaluža 2021, p. 5). This form of optimisation of content can be traced back to 2009, when Google enabled personalised searches based on different data-points like location, searching history, etc. (Kaluža 2021, p. 3). Since then, recommender systems have garnered expert attention, primarily because of the fear that people will "end-up" caught in echo-chambers – interacting with like-minded people that can bolster violence and extremism and therefore divide the public sphere into homogenous and isolated informational clusters (Burri et al. 2020, p. 43).

Internet filter-bubbles are often seen as one part of the complex dynamic that shapes an individual's informational space. A study from the Netherlands found that there are no serious concerns that people in this country are at risk of being caught into a filter bubble (Möller et al. 2019). However, another study that questioned the limited potential of internet filter bubble as a concept, focused on the algorithmic content personalisation and curation effect on media diversity. The study found that certain categories of people like young people and elderly, are indeed at risk of being exposed predominantly to algorithmically selected content (Bodó et al. 2019, p. 219). These findings align with the previously mentioned study in Bosnia and Herzegovina (see: Chapter I. Introduction) that illustrates how young people in the country are also more likely to consume content recommended by algorithms (Hodžić 2019, p. 32, 34).

[7] See here explanation: https://bit.ly/3OuTZbr.

The problem that the internet filter bubble concept seeks to capture and explain goes beyond a dualist understanding of the interactions between users and algorithms. The information consumption process is multi-dimensional – people use a range of resources to access information and actively search for content outside platforms and digital technologies. In addition to algorithms, there are other external socio-political and economic factors that curtail an individual's informational landscape (Bodó et al. 2019, p. 218; Mazzoli & Tambini 2020, p. 29). These arguments do not dismiss the power of algorithmic personalisation and recommendations systems governing content curation processes on the global level. The concerns tied to these processes remain legitimate because it is hard to foresee the kinds of consequences they will have on individuals and societies and their participation in the public sphere (Kaluža 2021, p. 14).

From the perspective of audio-video sharing and streaming platforms, these providers rely predominantly on recommendation systems to "rate" and "predict" preferences of individuals before "delivering" optimised audio-video content (Burri et al. 2020, p. 23). Contrary to news recommendation systems that are characterised by the influx and volume of information – thus requiring a different set of algorithmic-selection metrics – the recommendation systems deployed by AVMs providers essentially help users find relevant content without much effort. Thus, the role of informational abundance and diversity plays out differently in these two recommending domains (Bernstein et al. 2020, p. 49). However, the functional AI logic that drives both of these domains is more similar than it is not.

4.3. Hate speech and AI

The increasing deployment of automated systems to "police" speech and remove hateful and harmful content has proven to be a delicate process. Because AI systems are seen as context blind, they are unable to distinguish between sarcasm and hate speech. In addition, automated recommender systems are agile and constantly tweaked, and so is the nature of languages and online interactions (Yeung 2019, p. 29). Essentially, the problem boils down to the outcomes these automated systems produce. While at times they take down legitimate content (referred to as false negative), they also fail to remove "actual" hate-speech (false positive) (Bukovska 2020, p. 56). Keeping in mind the global application of these content removal systems, their unreliability illustrates the scale and their potential to limit individual freedom of expression online (Yeung 2019, p. 30; see also: Wagner et al. 2018, p. 17; OSCE SAIFE 2021, p. 17).

Besides these tech-related difficulties, there is still no universal definition of hate-speech and existing understandings tend to be interpreted and understood differently in different parts of the world. Furthermore, problems entailed by the use of the automated systems for removal of hateful content (e.g., Google Perspective) fail to address additional layers of local contextual complexities that are rarely taken into consideration in the process of developing these automated systems. Finally, responses to widespread hate speech and violence in digital space should not only be technology-led, but they also require inter-sectoral collaboration and multi-stakeholder strategic responses, including MIL interventions.

Algorithms and character assassination

For post-conflict societies, like the society of Bosnia and Herzegovina, hate speech and discriminatory practices profoundly colour public debate and peace narratives. The recent online attack – also referred to as "character assassination" (Turčilo et al. (forthcoming)) –

of two prominent Bosnian women and critical thinkers (Safetyofjournalists.net, 2021) has unlocked the negative potential of social media and online spaces, ending up in the off-line threats and harassment. The problem is not only hate speech and its impact on individuals, but also social media's infrastructure and AI logic that enables these problematic forms of speech and online behaviour to spread. Thus, the digital power of large technology companies has unleashed new forms of orchestrated harms and public distortions and so far, these companies are not showing much interest in addressing widespread and multiple forms of hate speech targeting individuals, ethnic groups, and critical thinkers in Bosnia and Herzegovina.

4.4. Deepfakes

Deepfakes are best described as AI generated videos and audios that create "alternative realities" or "synthetic media", making someone appear to be saying or doing things they never said or did. In their simplest form, deepfakes are created through computer programming and large data sets that are "fed by" images and audio of the person or object that deepfake should imitate (Burri et al. 2020, p. 147). Because of their potential for exploitation and abuse, they introduce numerous harms – from creating a potential swarm of misinformation or disinformation related media to generating online gender-based harassment and harm against individuals.

DeepNude

DeepNude has been discovered to swap women's faces into pornographic materials. This kind of harm created by deepfakes is labelled as a form of sexual extortion (Hao 2021). Deepfakes can be used to create non-consensual image sharing (sometimes colloquially referred to as revenge porn) of real people and can be used to trick people into believing that video content is an accurate representation (Burri et al. 2020, p. 147) .

The Tom Cruise Impersonator

In a viral Twitter and TikTok video, a person who is a Tom Cruise impersonator, used deepfake technology to turn his face into a nearly perfect mirror image of the actor Tom Cruise. This video was widely shared as it was believed to be the real actor Tom Cruise but was debunked as a deepfake a few hours later (Metz 2021). In addition, there are public consumer apps that allow for easy "face swapping" using deepfake technology (Zucconi 2018). From these multiple examples, we can see the harms of deepfake technology being utilised to generate misinformation and disinformation, along with impersonations of real individuals and violating users' privacy and agency.

4.5. Media freedom perspectives

Through a digital power lens, the following paragraphs look at the complex relationship between large technology companies and media freedom in the context of Bosnia and Herzegovina. Characterised by the lack of media pluralism, the media landscape of Bosnia and Herzegovina is shaped by a relatively large number of media outlets sharing a small advertising market and which are becoming increasingly dependent on social media to reach their audience and turn profit (Madoleva 2021; Kostić (forthcoming)). In fact, many media outlets are completely dependent on social media to disseminate their content, thus positioning social

media platforms as a powerful, yet largely "invisible", actor in the media landscape. This "invisibility" in reality means that for "economic and ideological reasons, search engines and social media companies have sought to remain content agnostic in terms of filtering out messages that could create public harms" (Donovan & Boyd 2019, p. 6).

4.5.1. Media pluralism. Journalism is an everchanging practice and media outlets are still struggling to sustain their societal relevance and public trust and strengthen their position and negotiation power vis-à-vis large technology companies. Consequently, media pluralism – the plurality of sources, content and exposure (Mazzoli & Tambini 2020, p. 40) – is radically reconfigured. This transformation means that these companies control the news dissemination and moderation processes, resulting in the so-called "lock-in" phenomenon (Burri et al. 2020, p.47) that takes negotiation power away from media actors on a global scale.

In addition, content moderation practises are not only non-transparent and exclude media agencies, but they are also dynamic and agile – constantly tweaking the opportunities for media outlets to reach the audience and in turn decreasing their revenue. In Bosnia and Herzegovina, where the media industry faces a dire financial situation and independent media outlets are also exposed to additional layers of political pressure and scrutiny (EC 2021, p. 23), the lack of revenue via social media infrastructure is a serious "blow" to their financial stability and media pluralism.

Democracy, a social media laboratory

In 2017, Facebook restructured its Newsfeed algorithms to increase the visibility of content shared and created by friends and advertisements. This form of reconfiguration made content produced from small independent news periodicals less visible to users in Slovakia, Sri Lanka, and Serbia (Hern 2017). KRIK's editor-in-chief, Stevan Dojčinović, in a New York Time op-ed article explains how large technology companies treat small democratic countries as laboratories and states: "[by] picking small countries with shaky democratic institutions to be experimental subjects, it is showing a cynical lack of concern for how its decisions affect the most vulnerable" (Dojčinović 2017). Ultimately, our democracies[8] have become a playing field and testing zone for social media companies.

4.5.2. Quality journalism. Through the embodiment of tools such as share and like buttons, media organisations – regardless of their size, impact, geographical origin and position in society – increasingly adopt the logic of social media companies. Consequently, we are witnessing the increased platformisation of the internet: "the rise of the platform as the dominant infrastructural and economic model of the social web and the consequences of the expansion of social media platforms into other spaces online" (Helmond 2015, p. 12). In this environment, media organisations exploit and benefit from these intrusive systems in pretty much the same way as large technology companies (e.g., by employing targeted advertising, sponsored content, or collecting and selling user data). Led by commercial interests, these media outlets produce content that co-modifies and seeks to capture the attention of an individual (e.g., "clickbait" or inflammatory content, see: Mattu et al. 2021; Donovan & Boyd 2019, p. 3, 7). Media organisations are implicitly incentivised to produce such content and sacrifice professional journalistic practises to ensure their financial viability (Mazzoli & Tambini 2020, p. 41). As journalist and editor Janus Rose, in an interview for the OSCE-funded project Explainable AI underscores: "[the] logic of platforms fundamentally runs contrary to

[8] See visual representation of the backdoor processes: Share Lab. (2016). Immaterial Labour and Data harvesting - Facebook Algorithmic Factory.

what we have aimed to produce as good quality journalistic content, which is sometimes not entirely based on what people want to read" (Explainable AI 2020).

Elections and troll farms

As previously noted, a functioning democracy requires open, vibrant, and unhindered public debate and free circulation of information. In countries where civic and media spaces are shrinking, the ability to sustain their democracy has become dire, as the following example shows.

One month before President Daniel Ortega ran for re-election in 2021, Facebook removed over 1,000 Facebook and Instagram accounts in an attempt to eradicate a troll farm – a clustered network of fake social media accounts coordinated to manipulate public discourse. According to Facebook, this troll farm was primarily operated by the Nicaraguan government and the country's ruling party, the Sandinista National Liberation Front (FSLN). It became active in April 2018 after student-led protests broke out against the government to discredit the protesters and increase pro-government content across multiple social media platforms: Facebook, TikTok, Twitter, YouTube, and Telegram. It should be noted that the Nicaragua government-linked troll farm is not an isolated event, but an increasing trend. In 2021, Facebook removed several other government-linked social media accounts from Ethiopia, Uganda, Sudan, Thailand, and Azerbaijan (Culliford 2021).

In 2020, Twitter removed around two thousand bot (fake) accounts in Serbia (The Guardian 2020). In 2019, similar manipulative activities were registered during the election period in Bosnia and Herzegovina (Cvjetićanin 2019, p. 7,22). Additionally, Wall Street Journal uncovered how Facebook has whitelists and blacklists for content, profiles, and pages (O'Neil 2021) which also indicates companies' gatekeeping practises and their effect on what we see online.

In this power game, social media conglomerates are the gold medal recipients. Their profit revenue continues to rise as the financial viability of news organisations becomes increasingly dependent on the services they supply, all meanwhile remaining apathetic towards citizen needs and agnostic towards the content they curate. Countries with complex media freedom situations, like Bosnia and Herzegovina, clearly experience the brute force of this dynamic. In this country, media organisations operate under a double opacity business model. In other words, media outlets pay large technology companies to promote their often low-quality content to make them visible and appealing to advertisers. Therefore, this content is often saturated with misinformation and hate speech and thus colouring the public's trust in them and the quality of the country's public debate (Sokol 2021, p. 20; Sokol 2020, p. 7; Cvjetićanin 2019, p. 22).

4.5.3. Media diversity. Like Newsfeed algorithms, news recommendation systems rely on user behaviour data and patterns to optimise and deliver personalised content. If a person is interested in reading content about sports, this person will likely be exposed to similar sports content, thus producing "a de facto reduction in the diversity of news offered to this consumer" (Bernstein et al. 2020, p. 52). By narrowing an individual's exposure and access to different points of view, values, and narratives, "content moderation practises set the agenda for the public's interests, interfere with democratic processes, and threaten informational pluralism and diversity" (Helberger 2019, p. 993). Thus, large technology companies dictate the structure of media freedom and often repress media diversity.

News aggregators clash with media diversity

In the current media environment, individuals are becoming increasingly dependent on news aggregators that collect and assemble news from a variety of sources in one place (OSCE RFoM 2021, p. 66; Newman et al. 2020, p. 11; Bernstein et al. 2020, p. 47). Although news aggregators systems utilise both human and algorithmic selection models, a recent study revealed that Apple Aggregator in the News Top Stories, which are curated by humans, offers more diverse content than the platform's algorithmically selected Trending Stories (Bandy & Diakopoulos 2020, p. 43). In addition to media diversity, news aggregators also impact the media landscape in two ways: aggregators tend to keep individuals engaged within their algorithmic news selection, thus limiting their use of other news content (also known as the "substitution effect") to the detriment of media financial viability and revenue-flow. However, the news aggregators have also provided more visibility to the less "popular" media outlets through the so-called "market-expansion effect" (OSCE SAIFE 2021, p. 67).

4.6. Interim conclusion

Overall, the issues outlined above illustrate the depth to which profit-driven AI logic and algorithmic systems are ingrained in our lives. While rooted in automation, the concerns highlighted above move beyond automation. Couldry and Mejias in their book, "The costs of connection. How Data Is Colonizing Human Life and Appropriating It for Capitalism" reveal that the problem is not one platform or one particular technology, but the interlocking combination of six key factors: existing (digital) infrastructure for extraction; an emerging order (meaning data practises that infuse humans with this infrastructure); an economic system built on the infrastructure and order; rationality that provides meaning to these practises; a new model of knowledge that reconfigures the world and "all there is to be known about human life" (Couldry & Mejias 2019, p. 192). From a macro perspective, these six factors illustrate how large technology companies are able to harness "opinion power" or the power to influence the "individual and public opinion formation" process (Helberger 2020, p. 846). In sum, "these platforms change the very structure and balance of the media market, and thereby directly and permanently impact the pluralistic public sphere" (ibid.). Opinion power inevitably shapes our freedom of expression (Burri et al. 2020, p. 62) because the extent to which we are able to practice freedom of expression is dependent on the kind of information we are exposed to and at what scale.

V. Regulatory and policy landscape: a brief overview

The sections presented above illustrate how algorithmic and data-driven systems shape a person's ability to receive and impart information in today's converging and 'platformasing' environment, therefore, raising the question: *who is responsible when AI impacts freedom of expression and media freedom?* To answer this question, this section briefly maps out AI related policy frameworks through the lens of the Council of Europe and the European Union. It should be noted that there is no binding regulatory framework addressing the conundrums of automated content moderation. The documents outlined below derive from the European Convention of Human Rights Article 10, which guarantees the protection of freedom of expression and covers all forms of expression, including through algorithmic systems and other communication technologies.

Before the widespread use of AI, the Council of Europe was involved in a number of issues concerning the emerging digital environment such as the role and responsibility of internet intermediaries and the relevance of the internet and new media in the era of technological convergence.[9] As AI systems became more prevalent, the Convention for the protection of individuals with regard to the automatic processing of individual data (hereafter Convention 108) was updated to focus on protecting individuals and their information in a Big Data world (T-PD(2017)01). These guidelines emphasise the relevance of human intervention in automated processes and informed user consent. Along the same lines, the Guidelines to Convention 108 on Artificial Intelligence and Data Protection (T-PD(2019)01) highlight how AI data processing practises need to align with general data protection principles and underline the relevance of informed consent and user agency (see: T-PD(2019)01, Part I). These guidelines also propose risk impact assessments to evaluate "the possible adverse consequences of AI applications on human rights and fundamental freedoms" (see: T-PD(2019)01, Part II) to ensure human intervention in instances of AI related human rights infringements (see: T-PD(2019)01, Part III). Additionally, the Recommendation of the Committee of Ministers to member States on Big Data for culture, literacy, and democracy (CM/Rec(2017)8) reaffirms the relevance of human rights approach (see: CM/Rec(2017)8, Preamble) and multi-stakeholder dialogue (para. 3) while pointing out the relevance of user's agency and control over algorithmic decision-making processes that "predict cultural attributes, preferences, and behaviour" (see: CM/Rec(2017)8, Preamble).

While the policies outlined above highlight the need for human rights-centred interventions, and underline the challenges AI and algorithmic systems pose, the Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes focuses on AI-driven misuse and their societal impact. This document proposes that safeguards against algorithmic manipulation and persuasion should go beyond data protection frameworks to "address the significant impacts of the targeted use of data on societies and on the exercise of human rights more broadly" (CM Decl(13/02/2019)1, 9.b) and suggest a clear delineation "between forms of permissible persuasion and unacceptable manipulation" (ibid., 9.c). In 2020 the Parliamentary Assembly of the Council of Europe (PACE/2341) adopted a resolution that highlights manipulative practises that may weaken and disrupt democracy (interfering in an electoral process, political micro-targeting, the amplification of propaganda, and polarisation and erosion of critical thinking (para. 14.4) and notes the relevance of cooperation among various stakeholders to identify "a set of commonly accepted principles on how to respond to concerns related to AI use" (para. 9).

Among these existing documents, the most comprehensive guidelines are set in the Recommendation CM/Rec(2020)1 of the Committee of Ministers on the human rights impacts of algorithmic systems. This policy proposes a human rights centred approach that refrains state and private actors from undertaking practises (including the development and use of algorithmic systems) that limit or impact human rights. To ensure these actors' compliance, they propose a supervisory and independent state body to "investigate, oversee and coordinate compliance with their relevant legislative and regulatory framework, in line with this recommendation" (para. 4) and engage in regular multi-stakeholder consultations, including

[9] See, for example: Recommendation (CM/Rec(2018)7) of the Committee of Ministers to member States on guidelines to respect, protect and fulfil the rights of the child in the digital environment.; Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries.; Parliamentary Assembly of the Council of Europe (PACE) Recommendation 2102(2017) on Technological convergence, artificial intelligence and human rights.

public educational and cultural institutions to ensure that "design, development and ongoing deployment of algorithmic systems are comprehensively monitored, debated and addressed" (para. 5).

The CoE's Committee of Ministers established an Ad Hoc Committee on Artificial Intelligence (CAHAI) essentially with the aim of exploring if specific regulation of AI systems in relation to human rights, democracy and the rule of law is necessary and how this could be done. Adopted in December 2020, the CAHAI's Feasibility Study introduces the notion that the risks to human rights should be perceived through a "socio-technical" AI lens to understand the values of the people who use and build them (Leslie et al. 2021, p. 13). CAHAI proposed the following nine key principles that underline its regulatory framework: human dignity, human freedom and autonomy, prevention of harm, non-discrimination, gender equality, fairness and diversity, transparency and explainability of AI systems, data protection and the right to privacy, accountability and responsibility, democracy and rule of law (ibid., p. 19). In December 2021, CAHAI adopted and submitted to the Committee of Ministers a document on the "Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law", which echoes some of the proposed regulatory elements introduced in its Feasibility Study. Similar to those mentioned in the PACE Resolution 2341, – such as election manipulation and profiling (Team AI Regulation 2021), – these proposed regulatory elements, inter alia, highlight the importance of addressing AI's impact on public opinion and its potential chilling effect on public participation.

In addition to these CoE regulatory initiatives, the EU has also initiated a number of regulatory interventions. By implementing these interventions, the EU aims to support technological innovation across all market actors (including large technology companies) while simultaneously addressing the potential harms they pose (Daly et al. 2021). This is evident in the 2020 White Paper on Artificial Intelligence: a European approach to excellence and trust, published by the European Commission. The frameworks they outlined in this paper to address human rights centred risks were followed by a number of sector specific regulatory instruments – among them is the drafted Digital Services Act (DSA).[10]

From a freedom of expression and media freedom perspective, the DSA is the most comprehensive and first EU-wide regulatory instrument to tackle the digital power of large technology companies. This act proposes a set of different obligations (e.g., mandatory auditing, transparency rules, access for researchers, etc.) for internet intermediaries divided into several categories, including "very large online platforms" that is to say large technology companies, as we referred to them (OSCE SAIFE 2021). To ensure a more level playing field among the actors of this digital market, the EU proposed the draft Digital Markets Act (DMA).[11] In addition to these policy regulations mechanisms, more topical instruments were adopted to target the spread of hate speech[12] and disinformation.[13] These topical instruments embrace multi-stakeholder and multidisciplinary models, while specifically involving large technology companies, like Facebook, Google and TikTok in problems related to disinformation (Burri et al. 2020, p. 40).

[10] European Commission. Proposal for a Regulation Of The European Parliament And Of The Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. COM/2020/825 final.
[11] European Commission. Proposal for a Regulation Of The European Parliament And Of The Council on contestable and fair markets in the digital sector (Digital Markets Act). COM/2020/842 final.
[12] European Commission. (2016). The EU Code of conduct on countering illegal hate speech online. (Adopted on 16 May 2016).
[13] European Commission. (2018). Code of Practice on Disinformation.

In mapping out AI related policy frameworks through the lens of the CoE and EU, the documents outlined above illustrate two main points: firstly, the increasing consensus that AI and algorithms pose a challenge to a range of human rights. To respond to these challenges, a human rights centred approach – a set of norms and values that protect human rights, democratic principles, and rule of law – emerges as a salient aspect of the regulatory design, development, deployment, and use of AI and algorithmic systems (PACE 2020, para. 11; CM/Rec(2020)1). The second point emphasises the responsibility, relevance, and engagement of multi-stakeholder and multidisciplinary perspectives in the policy design processes (CoE MSI-AUT(2018)06, para. 5; CM/Rec(2020)1, para. 5. B.1.3; Leslie et al. 2021, p. 31; Bernstein et al. 2020, p. 48).

The implementation of the efforts of these regulations further illustrates a growing momentum in a need to tackle the digital power large technologies harness. As demonstrated throughout this study, AI and algorithmic systems – these companies disempower individual agency, silence and deter minority groups, and increase societal tensions. The analysis of the regulatory policies outlined identify the normative framework that needs to be set forth to reduce the harms of automated technologies. The following section further discusses this framework in relation to MIL interventions.

5.1. Media and information literacy (MIL) and AI: the country perspective

As noted by the Council of Europe, MIL interventions are focal instruments in implementing human rights, democracy, and rule of law safeguards. Conceptually, MIL is constantly shifting to adhere to new socio-technological challenges (Chapman & Oermann 2020, p. 10). Therefore, as a framework it offers unlimited potential to engage individuals and societies. The documents analysed above recognise this potential of critical media, information, and digital literacy, for example, to understand the relevance of informed consent and ramifications of algorithmic systems MIL interventions should be designed as educational competences (CM/Rec(2020)1, para. 7; CM/Rec(2017)8; CoE MSI-AUT(2018)06, para. 7; T-PD(2017)01, para. 9). In addition, to minimise our exposure to AI related threats "[...]public awareness should be enhanced of the fact that algorithmic tools are widely used for commercial purposes and, increasingly, for political reasons, as well as for ambitions of anti- or undemocratic power gain, warfare, or to inflict direct harm " (CM Decl(13/02/2019)1, para. 9.e).

Against this backdrop, future MIL interventions should not only address technology-related aspects of AI and algorithmic systems but also:

- values, goals, and harms it poses to individuals and groups in a form of societal cohesion and peace processes;
- data-exploitation practises and business models of large technology companies;
- local contexts, local harms and opportunities it takes away from people.

In this way, MIL interventions introduce the complex and layered world of digital power and AI's impact on our daily lives. When implementing these interventions, particular attention must be placed on the elderly (over 65+) population, adults without formal or very low level of education, home caretakers, and anyone else who is not actively using or interacting with these digital technologies (every nine out of ten people in Bosnia and Herzegovina, see: CoE 2021, p. 5, 55, 65). To that end, the future tailored and targeted MIL interventions that address a plethora of related concerns (verification of resources, awareness of the widespread

use of AI and algorithms, content moderation and curation of personal informational spaces, ad-targeting and profiling, (see: Annex A) ultimately require genuine engagement and collaboration of different state and non-state actors.

VI. Conclusion

As illustrated throughout this study, the risks AI and algorithmic systems pose are erratic, volatile, and ever evolving. We are witnessing the radical reconfiguration that enables large technologies companies to harness digital power throughout the digital globe. In holding this power, the systems and technologies these companies deploy appropriate and exploit all aspects of human life. To protect our human rights, we must situate these companies and their socio-technical systems within a wider MIL framework to expose the values, goals, and agencies embedded within them. As a steppingstone in implementing these safeguards, this study explores MIL interventions as a framework for understanding the harms this multidimensional network of power poses on our autonomy and human rights while emphasising the relevance of a human rights and multi-stakeholder approach as not only a responsibility underpinning but also as a building block of MIL frameworks.

In this context, holistic and critical MIL interventions emerge as societal drivers that enable individuals to comprehend the scale and complexities of the harms these technologies pose and thus promote the ability to wisely engage with the services large technology companies supply. We are currently experiencing a transitional paradigmatic period. As we become increasingly dependent on AI and algorithmic technologies, a new communication and digital order is emerging. Right now, it is hard to foresee how AI and algorithms will continue to impact our lives and our freedom of expression. Therefore, studies, similar to this one, should be treated as an ongoing and iterative process that needs to be continuously updated.

Annex A

MIL Resources

The tables include useful resources that are organised in the following order:

1. Table 1:AI literacy
2. Table 2: digital safety toolbox and ways to increase user's agency and control over data
3. Table 3: resources on verification of credibility of digital content
4. Table 4: case-studies and additional illustrations

Table 1:AI Literacy

Algorithmic literacy, Kids Code Jeuness, CCUNESCO and UNESCO
A collection of critical essay and videos ,C.Sinders and B.Kostic, OSCE
Trips and tricks, learn more about AI,C.Sinders and B.Kostic, OSCE
Digital resilience, handbook for teachers, CoE
AI glossary ,CoE
Digital school for children, Digital School Slovenia
AI literacy resources, CoE data visualisation of AI initiatives
AI Literacy, Net Literacy
AI Literacy 101, Schouten, Towards Data Science
What to Read: A Biassed Guide to AI Literacy for the Beginner, P. Agre, MIT Libraries

Table 2: Digital safety tools

Safety tool-boxes, Front Line Defenders and Tactical Tech
Toolkit, Share Foundation (available in BHS)
Privacy protection tools and a video explainer, Share Foundation (available in BHS)
EFF Tools to protect privacy, Electronic Frontier Foundation
XYZ, gender and digital safety, Tactical Tech
Data resilience, Our Data Bodies
Totem project, courses, Free Press Unlimited

Table 3: Verification resources

A Beginner's Guide to Social Media Verification, Bellingcat
Insights and Recommendations for AI and Media Integrity, Partnership on AI.
Understanding and countering deepfakes, Witness
Analysis of browser extensions to flag suspicious content, Reuters Institute
Facebook split screen, Markup

Table 4: Case-studies and illustrations

A certain kind of doom scrolling, Caroline Sinders
Can data die?, The Pudding
Scary side of reality, Deepfakes, MIT Technology Review
Human rights violations, case studies and resources, Ranking Digital Rights
AI human factory and Facebook pyramid, Share Foundation
Meta-verse, Basic information, Washington post
Data Is power, Privacy International
How to make sure you don't take personalization too far, Harvard Business Review
Understanding difference between AI and Machine Learning, Microsoft

Annex B:

A glossary of AI related terms

Advertising tracking (ad-tracking): a practice of gathering personal insights regarding an individual's behaviours and preference to develop targeted and personalised advertisements.

Algorithms: a set of human-designed instructions with encoded procedures for transforming input data into the desired output, based on specific calculations.

Algorithmic harms: amplification and reinforcement of societal inequalities and discriminatory practices within the digital (hybrid) space that further marginalise disenfranchised communities and post-conflict societies (e.g., intersectional discrimination, amplification censorship, informational gaps, a large-scale information manipulation).

Artificial Intelligence (AI): a set of sciences, theories and techniques whose purpose is to reproduce by a machine the cognitive abilities of a human being.

AI bias: AI is written, created, and coded by people. The computational code and datasets that fuel algorithms, which in turn are the building blocks for AI, reflect and refract human values, biases, wants, and desires.

Automated decision-making systems: technological processes that augment and replace human decision-making processes with computer processors to answer different questions such as classification (e.g., female-male-non-binary) or assess score like in the cases of creditworthiness, risks crime occurrences.

Content moderation: processes of content monitoring, assessment, selection and distribution based on profiling and surveillance practise.

Cookies: a file created in the browser's directory to form a unique link between the individual and the site visited. All the actions performed on this site are stored within these small files.

Data repurposing: surveillance and data harvesting of individual personal and non-personal data across many different digital spaces and platforms for a specific purpose that is in the later stage recycled and used for different purposes.

Deepfakes: AI generated videos and audios that create 'alternative realities' or 'synthetic media', making someone appear to be saying or doing things they never said or did.

Digital power of AI: the concentration of data, information, and influence instilled in a small number of tech-based companies coupled with the increasing dependence of individuals, societies, and institutions on the provision of their services and the lack of effective democratic oversight.

Echo chamber: An environment in which a person is only exposed to beliefs and opinions that align with their own.

Impact of AI on people's decision making: AI and data-driven processes are deeply ingrained in and shape the context in which individuals make their decisions and choices, thus affecting their overall individual autonomy.

Internet bubble: Recommendation systems that expose people to similar homogeneous content that aligns with their interests and preferences.

Machine learning: type of (semi) autonomous learning and a developmental AI technique designed to improve the quality of automated decision making, by recognizing patterns and "regularities" to carry out certain tasks independent of human intervention.

Human in the loop: the need for human intervention within these automated processes to carry out oversight and corrective functions to ensure that individuals' human rights are respected and protected.

Platformisation of the internet: the rise of the platform as the dominant infrastructural and economic model of the social web and the consequences of the expansion of social media platforms into other spaces online.

Profiling: inferring personal details and behavioural patterns from seemingly uninteresting data through AI technologies to produce digital profiles that are used to inform future decision individuals make (e.g., for ad-tracking).

Responsible AI: the field that offers guidelines for the design, development, and deployment of AI that align with ethical and human rights standards.

Society in the loop: system that embeds values and interests of different societal groups, ethical and human right principles, and participatory opportunities in the design, development, implementation, and accessibility of the redress mechanisms of the automated systems for decision making.

User's control: the extent to which an individual can or cannot make decisions in regard to the use of their personal data and digital behaviour.

References

Legal and policy texts

Council of Europe. (1950). The European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols No. 11 and 14 (ETS No 5, 04/11/1950).

Council of Europe. (1997). Recommendation (R(97)20) of the Committee of Ministers to member states on "Hate Speech"

Council of Europe. (1981). Convention for the protection of individuals with regard to automatic processing of personal data (ETS No. 108, 28.01.1981).

Council of Europe. (2017). Guidelines to Convention 108 on the protection of individuals with regard to the processing of personal data in a world of Big Data. (T-PD(2017)01).

Council of Europe. (2017). Recommendation of the Committee of Ministers to member States on Big Data for culture, literacy and democracy. (CM/Rec(2017)8).

Council of Europe. (2018). Draft Recommendation of the Committee of Ministers to member States on the human rights impacts of algorithmic systems. (MSI-AUT(2018)06rev3).

Council of Europe. (2018). Recommendation (CM/Rec(2018)7) of the Committee of Ministers to member States on guidelines to respect, protect and fulfil the rights of the child in the digital environment.

Council of Europe. (2018). Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries.

Council of Europe. (2019). Declaration (CM Decl(13/02/2019)1) by the Committee of Ministers on the manipulative capabilities of algorithmic processes.

Council of Europe. (2019). Guidelines to Convention 108 on Artificial Intelligence and data protection. (T-PD(2019)01).

Council of Europe. (2020). Recommendation (CM/Rec(2020)1) of the Committee of Ministers to member Stateson the human rights impacts of algorithmic systems.

Council of Europe. (2020). PACE. Need for democratic governance of artificial intelligence. Doc. 15150. 24 September 2020

European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council 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 (General Data Protection Regulation). (Adopted on 27 April 2016).

European Commission. (2016). The EU Code of conduct on countering illegal hate speech online. (Adopted on 16 May 2016).

European Commission. (2018). Code of Practice on Disinformation.

European Commission. (2020) Proposal for a Regulation Of The European Parliament And Of The Council on contestable and fair markets in the digital sector (Digital Markets Act). COM/2020/842 final.

European Commission. (2020). Proposal for a Regulation Of The European Parliament And Of The Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. COM/2020/825 final.

European Commission. (2020). White Paper on Artificial Intelligence: a European approach to excellence and trust.

Parliamentary Assembly of the Council of Europe (PACE) Recommendation 2102(2017) on Technological convergence, artificial intelligence and human rights.

Parliamentary Assembly of the Council of Europe (PACE). (2020) Need for democratic governance of artificial intelligence. Resolution PACE/2341.

Parliamentary Assembly of the Council of Europe (PACE). (2020). Need for democratic governance of artificial intelligence. Report PACE/15150.

Literature

Al-Kawaz, H., N, Clarke., Furnell, S., Li, F., Abdulrahman, A. (2018). Advanced facial recognition for digital forensics. In A. Jøsang (Ed.), *Proceedings of the 17th European Conference on Information Warfare and Security: ECCWS 2018*. Academic Conferences and Publishing International Limited.

Bandy, J., & Diakopoulos, N. (2020). Auditing news curation systems: A case study examining algorithmic and editorial logic in Apple News. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 36-47).

Bernstein, A,. De Vreese Claes., Helberger, N.,S,Wolfgang., Z,Katharina . (2020). Diversity, Fairness, and Data-Driven Personalization in (News) Recommender System, *Dagstuhl Manifestos*, Dagstuhl Perspectives Workshop Vol. 9, Issue 1.

Bansal, A., Agarwal, R., & Sharma, R. K. (2012). SVM based gender classification using iris images. In *2012 Fourth International Conference on Computational Intelligence and Communication Networks* (pp. 425-429).

Baujard, T., Tereszkiewicz, R., de Swarte, A., Tuovinen, T. (2019). Entering the new paradigm of artificial intelligence and series. Council of Europe and Eurimages.Bloch-Wehba, H. (2020). Automation in moderation. Cornell International Law Journal. *53*, 41.

Bodó, B., Helberger, N., Eskens, S., & Möller, J. (2019). Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital journalism*, *7*(2).

Bukovska, B. (2020). Spotlight on Artificial Intelligence and Freedom of Expression. OSCE, The Representative on the Freedom of the Media (RFoM).

Burri, M., Eskens, S., Farish, K.,Frosio, G.,Guidotti, R., Jääskeläinen, A., Pin, A.,Raižytė, J. (2020). Artificial intelligence in the audiovisual sector. European Audiovisual Observatory.

Causevic, A. & Sengupta, A. (2020). Whose Knowledge Is Online? Practices of Epistemic Justice for a Digital New Deal. IT For change. Čaušević et al. 2020

Chapman, M., & Oermann, M. (2020). Supporting Quality Journalism through Media and Information Literacy.Council of Europe. Strasbourg.

Cobbe, J. (2020). Algorithmic Censorship by Social Platforms: Power and Resistance. Philosophy & Technology.

Couldry, N., & Mejias, U. A. (2019). The costs of connection. How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford University Press. Redwood City.

Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI). (2020). Toward Regulation of AI systems: Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe's standards on human rights, democracy and the rule of law.

Council of Europe Ad Hoc Committee On Artificial Intelligence (CAHAI). (2020). Feasibility Study. CAHAI(2020)23.

Council of Europe Commissioner for Human Rights.. (2019). Unboxing Artificial Intelligence: 10 steps to protect Human Rights.

Council of Europe. (2020). Digital Resistance: An empowering handbook for teachers on how to support their students to recognise fake news and false information found in the online environment.

Council of Europe. (2021). Conference Conclusions:Human Rights in the Era of AI Europe as international Standard Setter for Artificial Intelligence.

Council of Europe. (2021). Media Habits and Attitudes: Study on Media Habits of Adults in Bosnia and Herzegovina.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. London: Yale University Press.

Cvjetićanin, T. (2019). Dezinformacije u online sferi: slučaj BiH [Disinformation in online space: Bosnia and Herzegovina case study]. CSO Zašto ne?.Sarajevo.

Gerlitz, C., & Helmond, A. (2013).The like economy: Social buttons and the data-intensive web. *New Media & Society* 15(8).

Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B., & Wang, W. W. (2021). AI, Governance and Ethics: Global Perspectives. In H-W. Micklitz, O. Pollicino, A.

Amnon, A. Simoncini, G. Sartor, & G. De Gregorio (Eds.), *Constitutional Challenges in the Algorithmic Society* Cambridge University Press.

Dixon, P. (2013). Congressional Testimony: What Information Do Data Brokers Have on Consumers?. World Privacy Forum.

Donovan, J., & Boyd, D. (2019). Stop the presses? Moving from strategic silence to strategic amplification in a networked media ecosystem. *American Behavioral Scientist*, 1–18. SAGE Publication.

European Commission Policy Department for External Politics. (2021). Mapping Fake News and Disinformation in the Western Balkans and Identifying Ways to Effectively Counter Them. Brussels.

European Commission. (2020). Commission staff working document, Bosnia and Herzegovina 2020 Report. Accompanying the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (SWD(2020) 350 final). Brussels.

European Digital Rights Initiative. (2020). DSA: Platform Regulation Done Right.

Fairhurst, M., Da Costa-Abreu, M., & Erbilek, M. (2015). Exploring gender prediction from iris biometrics. In *2015 International Conference of the Biometrics Special Interest Group (BIOSIG)* (pp. 1-11). IEEE.

Gillespie, T. (2014). The relevance of algorithms. *In Media technologies: Essays on communication, materiality, and society.* MIT Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*. SAGE.

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism, 7*(8), 993-1012.

Helberger, N. (2020). The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism*, *8*(6), 842-854.

Helberger, N., Eskens, S. J., van Drunen, M. Z., Bastian, M. B., & Möller, J. E. (2019). Implications of AI-driven tools in the media for freedom of expression. In *Artificial intelligence–Intelligent politics: Challenges and opportunities for media and democracy*.

Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society*.

Hodžić, S. (2019). Surfanje po tankom ledu:mladi, mediji i problematican sadrzaja [Surfing on the thin ice: youth, media and problematic content]. Media Center Sarajevo. Sarajevo.

Jenkins, H. (2006). Convergence Culture: Where Old and New Media Collide. New York. New York University Press.

Jørgensen, R. F. (2019). *Human rights in the age of* platforms. The MIT Press.

Kalluri, P. (2020). Don't ask if AI is good or fair, ask how it shifts power. World View. Nature Vol. 583. Springer.

Keller, D. (2021). Amplification and Its Discontents.Occasional Papers. Knight First Amendment Institute. Colombia University.

Kaluža, J. (2021). Habitual Generation of Filter Bubbles: Why is Algorithmic Personalisation Problematic for the Democratic Public Sphere?, Javnost - The Public.

Klonick, K. (2017). The New Governors: The People, Rules, and Processes Governing Online Speech. 131 Harv. L. Rev. 1598.

Kostić, B. (forthcoming). Report: Content moderation on Social Media and Map of Stakeholders in Bosnia and Herzegovina. ARTICLE 19.

Kostić, B. (2021). Veštačka inteligencija - Uticaj na slobodu izražavanja, medijske perspektive i regulatorni trendovi [Artificial Intelligence - Impact on freedom of expression, media perspectives, and regulathory trends] (2021) OSCE Serbia. Available at: https://www.osce.org/sr/mission-to-serbia/479672

Kuehlkamp, A., Becker, B., & Bowyer, K. (2017). Gender-from-iris or gender-from-mascara?. In *2017 IEEE Winter conference on applications of computer vision (WACV)*(pp. 1151-1159). IEEE.

Lagree, S., & Bowyer, K. W. (2011). Predicting ethnicity and gender from iris texture. In *2011 IEEE international conference on technologies for homeland security (Hst)* (pp. 440-445). IEEE.

Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M., Briggs, M. (2021). Artificial intelligence, human rights, democracy and the rule of law. Council of Europe.

Leerssen, P. (2020). The Soap Box as a Black Box: Regulating transparency in social media recommender systems. *European Journal of Law and Technology*, *11*(2).

Maréchal, N., & Biddle, E. R. (2020). It's not just the content, it's the business model: democracy's online speech challenge. *New America—Ranking Digital Rights, March*, *17*.

Mazzoli, M.E., & Tambini, D. (2020). Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe.

Möller, J., Helberger, N., & Makhortykh, M. (2019). Filter Bubbles in The Netherland. Hilversum: Commissariaat voor de Media.

Molnar, P. (2020). Technological Testing Grounds: Border tech is experimenting with people's lives. European Digital Rights (EDRi). Brussels.

Mozilla. (2021). YouTube Regrets. A crowdsourced investigation into YouTube's recommendation algorithm.

Newman, N., Fletcher, R., Schulz, A., Andı, S., and Nielsen, R.K. (2020). Reuters Institute Digital News Report 2020. Reuters Institute for the Study of Journalism. Oxford University.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

Onuoha, M. & D, Nucera. (2018). A People's Guide to AI : Artificial Intelligence. Open Society Foundations.

Organisation for Security and Cooperation in Europe Representative on Freedom of the Media (OSCE RFoM). (2021). A Policy Manual. Spotlight on Artificial Intelligence and Freedom of Expression # SAIFE.

Organisation for Security and Cooperation in Europe Representative on Freedom of the Media (OSCE RFoM). (2020). Non-paper on the impact of artificial intelligence on freedom of expression. #SAIFE.

Partnership on AI. (2020). The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity.

Pasquale, F. (2015). The Black Box Society:The Secret Algorithms That Control Money and Information. Harvard University Press. Cambridge Massachusetts.

Privacy International & Article 19. (2018). Privacy and Freedom of expression in the age of artificial intelligence. London.

Privacy International. (2017). Data is power: Profiling and Automated Decision-Making in GDPR. Privacy International. London.

Rahwan, I. (2017). Society in the Loop: Programming the Algorithmic Social Contract, Ethics of Information Technology.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-23.

Ranking Digital Rights. (2020). Human rights risk scenarios: Algorithms, machine learning and automated decision-making (Consultation Draft).

Sinders, C. (2021). Designing Against Dark Patterns.German Marchall Fund of the United States.

Sokol, A. (2020). Propaganda, Disinformation and hate models of media and communication in Bosnia and Herzegovina. SEENPM, Tirana, Peace Institute, Ljubljana and Foundation Mediacentar Sarajevo. Sarajevo.

Sokol, A. (2021). Hate narratives in the media and user-generated content. Media Center Sarajevo. Sarajevo.

Tapia, J. E., Perez, C. A., & Bowyer, K. W. (2014). Gender classification from iris images using fusion of uniform local binary patterns. In *European Conference on Computer Vision* (pp. 751-763). Springer International Publishing.

Tapia, J. E., Perez, C. A., & Bowyer, K. W. (2016). Gender classification from the same iris code used for recognition. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1760-1770.

Thomas, V., Chawla, N. V., Bowyer, K. W., and Flynn, P. J. (2007). Learning to predict gender from iris images. In *First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS).*

Turčilo, L. & Buljubašić, B. (forthcoming). Uništavanje reputacije na bh nacin:govor mržnje botova u online prostoru kao sredstvo sužavanja javnog prostora za alternativna mišljenja u Bosni i Hercegovini [Destruction of reputation in bh.way: bot hate speech in online space as a tool to reduce public space for alternative views in Bosnia and Herzegovina]. In *Conference Komentari, govor mržnje, dezinformacija i regulacija javne komunikacije. Agencija za elektronicke medije i Medijska istraživanja.* Zagreb.

UC Berkeley School of Law Human Rights Center Research Team.(2019). Memorandum on Artificial Intelligence and Child Rights.

UNESCO & EQUALS Skills Coalition. (2019). I'd blush if I could: closing gender divides in digital skills through education.

UNESCO. (2020). Artificial Intelligence and Inclusion: Compendium of Promising Initiatives. Mobile Learning Week 2020.

United Nations Human Rights Council (UNHRC). (2018) Report of the independent international fact-finding mission on Myanmar. (A/HRC/39/64).

Wagner, Ben, Wolfgang Schulz, Karmen Turk, Bertrand de la Chapelle, Julia Hörnle, Tanja Kersevan-Smokvina, Mathias C. Kettemann, Dörte Nieland, Arseny Nedyak, Pēteris Podvinskis, Thomas Schneider, Sophie Stalla-Bourdillon, and Dirk Voorhoof. (2018) Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications. DGI(2017)12. Council of Europe.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe.

Yeung, K. (2019). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe.

Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe.

News articles, blogs and op-eds

Ackerman, E. (2021). This Year Autonomous Truck will take on the road with no one on board. IEEE Spectrum.

Appen. (2019). Human in the loop Mechine Learning.

Association for the Advancement of Artificial Intelligence (AAAI).

BBC. (2019). Apple's 'sexist' credit card investigated by US regulator.

Bell, G. (2020). Touching the future. The Griffith Review.

Brandom, R. (2021). Social media vs. the world: What comes after deplatforming. The Verge.

Council of Europe AI Glossary

Constine, J. (2016). How Facebook News Feed Works. Techcrunch.

Culliford, E.(2021). Facebook says it removed troll farm run by Nicaraguan government. Swissinfo.ch.

Dojčinović, S. (2017). Hey, Mark Zuckerberg: My Democracy Isn't Your Laboratory. The New York Times.

Duhigg, C. (2012). How Companies Learn Your Secrets. The New York Times.

Europan Commission. (2019). The Digital Markets Act: ensuring fair and open digital markets.

European Citizens' Initiative (ECI). (2021). Reclaim your face.

Explainable AI. (2020). Speech and AI.

Feldstein, S. (2020). When It Comes to Digital Authoritarianism, China is a Challenge—But Not the Only Challenge. Carnegie Endowment for International Peace.

Gilman,M. (2020). AI algorithms intended to root out welfare fraud often end up punishing the poor instead. The Conversation.

Global Internet Forum to Counter Terrorism (GIFCT). Sharing Consortium Hashtag.

Google. Google Perspective.

Ghoshal,D.(2018). Mapped: The breathtaking global reach of Cambridge Analytica's parent company. Quartz.

Hao, K. (2021). A horrifying new AI app swaps women into porn videos with a click. MIT Review.

Harrison, S. (2020). Why Am I Paying $60 for That Bag of Rice on Amazon.com?. The Markup. (2020).

Hern, A. (2017). Facebook moving non-promoted posts out of news feed in trial. The Guardian.

Horwitz, J. & Seetharama, D. (2020). Facebook executives shut down efforts to make the site less divisive. The Wall Street Journal.

IBM Cloud Education. (2020). Artificial Intelligence (AI).

Jeremić, I. & Stojanovic, M. (2021). Facebook, Twitter Struggling in Fight against Balkan Content Violations, Balkan Insight, BIRN.

Kastrenakes, J. (2016). Microsoft made a chatbot that tweets like a teen. The Verge.

Leetaru,K. (2018). The Data Brokers So Powerful Even Facebook Bought Their Data - But They Got Me Wildly Wrong. The Forbs.

Leufer, D. (2021). Computers are binary, people are not: how AI systems undermine LGBTQ identity. Access now.

Lewis, P. & Mc.Cormick, E. (2018). How an ex-YouTube insider investigated its secret algorithm. The Guardian.

Madoleva, S. (2021). Media Freedom in Bosnia and Herzegovina. Konrad Adenauer Stiftung.

Mattu, S., Yin, L., Waller, A., & Keegan, J. (2021). How We Built a Facebook Inspector. The Markup.

Merrill, J. (2021). How Facebook's Ad System Lets Companies Talk Out of Both Sides of Their Mouths. The Markup.

Merrill, J. & Oremus, W. (2021) Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. The Washington Post.

Metz, R. (2021). How a deepfake Tom Cruise on TikTok turned into a very real AI company. CNN.

Newton, C. (2021). The tier list: how Facebook decides which countries need protection. The Verge.

Nikolic, I & Jeremic,I. (2020). 'Vox Populi': How Serbian Tabloids and Twitter Bots Joined Forces. Balkan Insight. BIRN.

O'Neil, C. (2021). Facebook's VIP 'Whitelist' Reveals Two Big Problems. Bloomberg.

Our Data Bodies.

Ohlheiser, A.(2019) The one word that lets politicians get away with breaking the rules on social media. Washington Post.

Pirkova,E. (2020). How the Digital Services Act could hack Big Tech's human rights problem. Access now.

Safetyofjournalists.net. (2021). The brutal campaign against Lejla Turcilo and Borka Rudic from Sarajevo must stop.

Statista.com. (2021). Facebook users in Bosnia and Herzegovina from September 2018 to July 2021.

Team AI Regulation. (2021). The Council of Europe's recommendation for a legal framework on AI. AI Regulation.

The Guardian. (2019). New York tenants fight as landlords embrace facial recognition cameras.

The Guardian. (2020). Twitter deletes 20,000 fake accounts linked to Saudi, Serbian and Egyptian governments.

The Markup. (2021). How Facebook's Ad System Lets Companies Talk Out of Both Sides of Their Mouths.

Zucconi, A. (2018). An Introduction to Deepfakes. AlanZucconi.com.

This study provides a critical insight into AI's current landscape and the challenges it poses on human rights. It should be noted that this study does not cover all relevant aspects of AI and human rights as many AI related topics fall outside its scope – such as sector specific use of AI, data protection concerns outside of AI and freedom of expression, as well as oversight and human rights assessment standards. Some of these aspects are mentioned as illustrative examples or supporting arguments, but the focus remains on articulating and mapping out the current expert and regulatory discussions through the lens of AI and freedom of expression.

www.coe.int/freedomofexpression

www.coe.int

The **Council of Europe** is the continent's leading human rights organisation. It comprises 46 member states, including all members of the European Union. All Council of Europe member states have signed up to the European Convention on Human Rights, a treaty designed to protect human rights, democracy and the rule of law. The European Court of Human Rights oversees the implementation of the Convention in the member states.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE