Public perception of algorithms in society: accounting for the algorithmic public opinion .

Algocount 2020-2022

founded by :

Fondazione CARIPLO

The Algorithmic Public Opinion: a Policy Overview

Urbano Reviglio, Università degli Studi di Milano

The **ALGOCOUNT** project seeks to contribute to concerns on the political and policy implications of the rise of algorithms. The diffusion of an "algorithmic public opinion" specifically regards the concern that algorithmic systems disrupt previous patterns of individual and public opinion formation. Few analyses explored this phenomenon throughout a multidisciplinary comprehensive policy lens. The **ALGOCOUNT** project seeks to address this gap by discussing innovative policy approaches and proposals to the challenges of the current algorithmic public opinion. Firstly, I introduce the main algorithms, concerns and actors involved in the algorithmic public opinion. Then, I discuss the main policy approaches to the systemic challenges of the most relevant algorithmic systems: recommender systems, disinformation, political microtargeting, content moderation and social bots. The goal of this report is to provide a critical, comprehensive and long-term oriented policy overview of the algorithmic public opinion.

Table of Contents

An Introduction to the Algorithmic Public Opinion 4

Concerns Related to the Algorithmic Public Opinion 10

Main Actors in the Algorithmic Public Opinion 18

The Governance of the Algorithmic Public Opinion 29

Conclusions 52

Executive Summary

- → The algorithmic public opinion comprises several algorithms which profoundly influence worldwide information flows - information production, distribution and consumption - and, therefore, intermediating – not determining – opinion formation both at the individual and collective level.
- → The most influencing algorithms in this context are arguably personalization algorithms that act as fundamental global gatekeepers. Yet, other relevant algorithms are the ones employed in content moderation (to detect and remove content), social bots (to manipulate content diffusion), and, among others, those who help to create and edit content.
- → More than decreased media diversity (e.g. filter bubble theory and echo chambers), the most concerning issues in the algorithmic public opinion are algorithmic censorship, algorithmic manipulation and, more generally, disinformation.
- → It is paramount to allow researchers to access wider datasets in a legal and transparent manner as well as to let them run on-platform experiments capable of detecting more reliable causal relationships. This is a condition *sine qua non* for collecting evidence for effective policies. Importantly, policies for the algorithmic public opinion can even increase the quality of public policies more generally.
- → The question of regulating the algorithmic public opinion is part of a complex regulatory system involving many different legislative acts and normative approaches. Policy-making still lacks a shared vocabulary or frameworks for approaching the incoming challenges and new models for digital governance will likely need to be developed. The algorithmic public opinion is in fact particularly challenging to policy-makers as it intersects with a transformation of media itself, due to media convergence, media globalization and the rise of global digital platforms.
- → It is desirable to foster citizens' active participation in the design and governance of digital platforms, rather than relying on top-down regulations by a technocratic elite. Similarly, a truly multistakeholder governance has to allow civil society a voice, "loud" enough to be heard. A truly multistakeholder governance approach in which civil society has finally a voice is necessary.
- → The role of ethics in mediating the regulatory process is essential. Despite widespread phenomena such as ethics washing and ethics bashing who threaten to limit its potential relevance and influence, ethical debates are necessary to develop better and more inclusive policies and guidelines and, therefore, deserve more attention from all the involved stakeholders.
- → As traditional news media is still struggling with the opportunities and challenges of digital societies as well as the emerging algorithmic public opinion, it is fundamental to understand its essential role in re-mediating the algorithmic public opinion. News quality, thus a well-informed public opinion, highly depends on

the ability of news media to remain politically and economically independent so as to provide high quality journalism.

→ US and European policymakers have mostly focused on the latest controversies and the biggest players. To tackle urgent policies we need to understand algorithms as expansive socio-technical phenomena that functions in many contexts and takes many forms. Expanding the scope and range of research is critical to developing sound policy. The range of contentious phenomena cannot be captured with sectorial approaches. Fundamentally, the largest, US-based platforms do not provide a reliable guide for the entire social media ecology.

→ For a sustainable algorithmic public opinion, it is needed a paradigm change. Admittedly, the main concerns and risks associated with the algorithmic public opinion are mainly rooted in the perverse dynamics arising from the current mainstream business model based on advertising. This model is at the heart of surveillance capitalism which is widely believed it needs structural reforms.

→ It is necessary to focus on policies to reinforce the quality and effectiveness of policies themselves. There are indeed structural concerns of lobbying pressures from a variety of actors, in primis big tech companies, to consistently undermine the most innovative and effective impetuses of new regulations. Similarly, there should be stricter rules on the phenomenon of revolving doors.

→ Finally, we should avoid naive expectations of the Internet as an ideal public sphere: there are always new challenges, and human and machines mistakes will inevitably raise further concerns. It is a never-ending challenge the governance of the algorithmic public opinion.

An Introduction to the Algorithmic Public Opinion

Algorithms are everywhere nowadays. From traffic lights to election machines, from stock markets to pacemakers: there is a wide range of algorithmic variability, complexity and relevance. In the media landscape, they are used in various contexts for different purposes; from the selection of personalized content to the detection and removal of fake news and harmful or copyrighted content, from the selection of news for journalists to the automated creation of news articles, from the usage of social bots to the creation of deepfakes. Algorithms are indeed becoming an essential intermediary of how information is produced, disseminated and consumed, and thus how public opinion is formed, giving rise to what we define as an "algorithmic public opinion" (cfr. D6).

What are algorithms and how they really work, however, is not easy to explain, if at all. In public discussions, the term "algorithms" has indeed become a 'sloppy shorthands, slang terms for the act of mistaking multipart complex systems for simple, singular ones' **(Bogost, 2015)**. Algorithms are more than one (neutral) technology. As Lessig (1999) famously pointed out, rules are also implicit in code – in all the algorithms that select, prioritize, remove and, ultimately, affect the public opinion. More broadly, an algorithm can be intended as "a socio-economic construct, that is, a technology that is embedded in organisations with their own goals, values and fundamental freedoms, and that mediate and impact interactions with the human/economic/social environment in which they are functioning" (**Helberger et al, 2019, p. 3**). There are, indeed, several human influences embedded into algorithms, such as criteria choices, training data, semantics, and interpretation. Algorithms need to be seen and analyzed as a complex assemblage of procedures, individuals and teams, ideology, datasets constantly changing and difficult to reduce to a singular, simple concept. To fully understand them and their consequences is therefore fundamental to understand both their technical complexity and variability on the one hand, and their fundamentally social embeddedness on the other.

Types of Algorithmic Systems

There are several algorithmic systems which influence the formation of public opinion. These can affect the news ecosystem on at least three levels: news production and distribution, individual users, the broader media landscape and society more generally. Algorithms exert power through decisions they make, for example, in prioritizing information (e.g. search engines), filtering news (e.g. Facebook, Youtube etc.), making classifications (e.g. Airbnb through the reputation system) or finding associations (e.g. Google Flu Trends). More generally, algorithms can automate several actions; from smart tools that assist journalists in producing their stories and fully automated production of news stories (so-called robot journalism) to the usage of social bots and the

creation of deepfakes for propaganda purposes: their role is various, pervasive and is increasingly institutionalized. In this policy review, I have identified four significant influential algorithmic systems: content recommendation, arguably the most influential one as it personalizes our information environment; automated content analysis tools (content moderation); social bots; and, finally, the ones employed in news media.

Content Recommendation

With the advent of the Internet, traditional models of information and news production, distribution and consumption have radically changed. In order to provide relevant information to users, the epochal transition from information scarcity to information abundance brought the need to balance user preferences and algorithmic delegation (**Thurman and Schifferes, 2012**). As such, user interests became more important than content quality or social significance (**DeVito, 2017**). Such development made personalization very appealing in our individualistic societies. By using online services, the audience can exert a greater control over news selection (in theory) and eventually, can focus on issues that they consider more relevant, which in turn empowers audiences and erodes the degree of editorial influence over the public's issue agenda. This is basically how personalization algorithms legitimized mainstream social media's gatekeeping role. The reality, however, is much more complicated and nuanced.

Everyday people and companies benefit from the personalization of recommender systems. For example, when Netflix recommends a movie, Youtube a video, Facebook a stream of information or potential friends, Spotify music and Amazon related products, a recommendation algorithm is what performs these tasks. Recommendation and personalization are entrenched; the functioning of social media sites is inextricably linked to the quality of their personalization algorithms. Our time and attention are indeed scarce resources whereas the quantity of potentially relevant information immense. Personalization systems therefore perform this fundamental role of knowledge management. They are both socio-technical systems that mediate and influence interactions with the economic and social environment as well as technologies of construction of the self that significantly contribute to affect in the longer term who we are and who we think we are.

Online news consumption is increasingly relevant in terms of public opinion's formation. According to existing research (**Newman, 2016**), almost 70 percent of online news users surveyed across 37 different markets worldwide identified distributed forms of discovery as their main way of accessing and finding news online, with search and social media being by far the most influential factors, followed by aggregators, email and mobile alerts. Their influence is substantial. Consider three of the most visited websites; on Facebook, the posts encountered by the average user everyday are circa 350 on about at least 1.500 potential posts (**Backstrom, 2013**). Thus, roughly 25% are algorithmically filtered and prioritized and 75% are ultimately hidden to individual users. On Youtube, algorithms already drive more than 70% of the time spent in

the video sharing platform and 90% of the 'related content' in the right bar is indeed personalized. On Netflix, around 80% of viewing hours come through recommendations and only around 20% from its search function.

(see → Box 1 *The Reductionism of Profiling Technologies*)

These recommender systems are generally used for personalization. Personalization is defined as "a form of user-to-system interactivity that uses a set of technological features to adapt the content, delivery, and arrangement of a communication to individual users' explicitly registered and/or implicitly determined preferences" (**Thurman et al., 2011, p.3**). It can in fact be explicit or implicit, that is, it can depend on user's requests and/or user's behavioural data (which is mostly created unknowingly and unwittingly). Personalization can thus be based on the individual autonomy of choice or on the algorithmic/platform delegation to infer one's personal preferences. Both implicit and explicit personalization increased dramatically in the last years, though many websites have acted to make passive forms of personalization the fastest growing forms (**Thurman, 2011**). This raises further concerns about unintended consequences and the reductionism of profiling technologies.

Personalization algorithms are also employed for microtargeting, that is predictive market segmentation, or even nanotargeting, namely delivering ads exclusively to a specific user (**González-Cabañas et al., 2021**). These are used for targeted political advertisements, so-called political microtargeting. These help political parties to identify the individual voters that it is most likely to convince and, at the same time, match its message to the specific interests and vulnerabilities of specific voters. These techniques promise to make propaganda even more tailored to individual voters, and more effective. So far, they are primarily used in the United States, but have recently gained popularity in European countries too (**Zuiderveen Borgesius et al., 2018**).

Content Moderation

Online moderation systems are methods to sort contributions that are irrelevant, obscene or illegal in order to ensure that the content complies with legal and regulatory requirements, site/community guidelines and user agreements. In recent years, there has been an expansion in research and investment in automated content analysis tools (see **Shenkman et al., 2021**). This has been further accelerated during the COVID-19 pandemic due to concerns about health risks. They help to solve the challenge of scale, in particular to inspect or remove content flagged for hate speech or other objectionable or problematic content. There are, essentially, two types of automated content analysis:

1. Matching models which are generally well-suited for analyzing known, existing images, audio, and video, particularly where the same content tends to be circulated repeatedly and
2. Predictive models that can be well-suited to analyzing content for

which ample and comprehensive training data is available. Examples can include whether multimedia contains clear nudity, blood, or discrete objects.

These algorithms are mostly designed to detect child sexual abuse material, terrorist propoganda, nudity, hate speech and copyrighted content. Notably, Facebook utilizes a largescale matching algorithm on "every image uploaded to Instagram and Facebook" to scan against an existing curated database of "misinformation," including COVID-19 misinformation. But there are various other usages; for instance, any website can use these algorithms to support or automate comment section moderation. Online social matchmaking services like OkCupid have utilized algorithms to scan for re-uploads of banned profiles. Amazon utilizes 'audio fingerprinting' to prevent mentions of the word "Alexa" in advertisements from mistakenly triggering Alexa devices and resulting in negative customer experiences. In essence, these algorithmic systems help to detect content with certain features to eventually remove them.

The overall impact of these predictive systems on the speech of platform users is still poorly understood. While there are many important and useful advances being made in the capabilities of machine learning techniques to analyze content, policymakers, technology companies, journalists, advocates, and other stakeholders need to understand and deal with the limitations of these algorithms. The risk is that they become a form of censorship and eventually they will lead to other detrimental impacts on human rights, and on the ability of platforms to function as spaces for discourse, communication, and interpersonal relation (see **Cobbe, 2020**).

Social Bots

Bots are automated computer software that perform tasks along a set of algorithms, and they are at work all over the Internet at varying levels of sophistication. According to one oft-cited estimate[1], over 37% of all Internet traffic is not human and is instead the work of bots designed for either good or bad purposes. At the same time, a study showed that 30% of users can be deceived by a bot. A team of researchers from the University of Southern Carolina and Indiana University released figures suggesting that between 9% and 15% of active Twitter accounts are bots, while higher percentages of politician's followers are also fake. They can eventually be used for propaganda: russian bots tweeted conspiracy theories at US president Donald Trump in an effort to get him to spread the stories through the media.

Social bots legitimate uses vary: crawler bots collect data for search engine optimization or market analysis; monitoring bots analyze website and system health; aggregator bots gather information and news from different sources; and chatbots simulate human conversation to provide automated customer support. **Gorwa and Guilbeault (2020)** identify six main bot types:

1 https://www.imperva.com/resources/resource-library/reports/2020-Bad-Bot-Report/.

1. Web Robots, crawlers and scrapers which download and index websites in bulk.
2. Chatbots, programs that approximate human speech and interact with humans directly through some sort of interface.
3. Spambots, bots that can be used to send spam en masse or perform DDoS attacks.
4. Sockpuppets and Trolls, bots which can be deployed by government employees or regular users trying to influence discussions, to fabricate reviews and post fake comments about products, people or institutions.
5. Cyborgs and Hybrid Accounts which are hybrid bots: bot-assisted human or human-assisted bot.
6. Last but not least Social Bots, bots that automatically produces content and interacts with humans on social media.

Social bots are likely the most influential one in the context of public opinion, as they are very common and have a wide variety of uses. The Department of Homeland Security describes them as programs that "can be used on social media platforms to do various useful and malicious tasks while simulating human behavior." Lutz Finger identified 5 uses for social bots relevant to the algorithmic public opinion: 1) foster fame: having an arbitrary number of (unrevealed) bots as (fake) followers can help simulate real success in a sort of self-fulfilling prophecy. 2) Spamming: having advertising bots in online chats is similar to email spam, but a lot more direct mischief: e.g. signing up an opponent with a lot of fake identities and spam the account or help others discover it to discreditize the opponent. 3) Limit free speech: important messages can be pushed out of sight by a deluge of automated bot messages. 4) Fishing: to fish passwords or other personal data. And finally 5) bias public opinion: influence trends by countless messages of similar content with different phrasings (e.g. hashtag flooding or tweetbombing, to make certain hashtag trending topic) or even create fake grass-roots movements (also called astroturfing, which is comes from 'astroturf' meaning 'fake grass', known in the online context as 'cyberturfing' **(Cobbe and Singh, 2019)**.

In essence, social bots attempt to 'game' the algorithm by inflating the 'reputation' of content and thus increase its likelihood of being recommended or its position in algorithmic content rankings. By posting content strategically and artificially inflating views, likes, shares, and other metrics, networks of bots can together shape the construction of online spaces. They can therefore artificially seed political messages in organic discussions, bring greater attention to stories (real or fake), and boost ideas (fringe or otherwise) into mainstream discussion.

Algorithms in News Media

Several innovative algorithms and 'AI' are increasingly used in the production of media news and journalism for various scopes, especially in computational journalism (see **Diakopoulos, 2014**). These generally help innovate the production and consumption of news and the

journalistic investigations as well as provide better services at lower costs. Here I summarize the main algorithms employed in news media and journalism providing concrete examples.

Automated production of news (robot journalism). The Press Association in the UK produces more than 30.000 local info per month using AI tools. In 2015, LeMonde used a technology called Syllabs during the departmental elections to write local articles on the results of the 36,000 municipalities and cantons affected by the elections. The Norwegian News Agency produced sports report 30 seconds after a football game. AI helps journalists find new angles to a story. For example, the INJECT project helps journalists scanning articles on a given topic and makes a proposal for a different angle by suggesting the story via different actors, submitting related cartoons or data, challenging by asking incidental questions, or producing facts cards on the topic.

Investigative journalism. For example, data driven technologies and AI could provide significant improvements to the in depth analysis of vast amounts of data of journalistic value such as the Panama Papers or other leaks. More generally, data mining and processing, and AI can help journalists to provide more accurate information and develop a more comprehensive understanding of reality. For example, Serelay is a start-up which allows verifying if a supposed location of a captured image is correct and check if no post-processing of this image took place.

Audience analytics. The British news outlet *The Guardian* has developed algorithms that allow journalists to understand how audiences are responding to content - whether the audiences are reading the texts thoroughly or engaging with it in a particular way. The Neue *Zürcher Zeitung* uses a personalized and dynamic paywall with 150 criteria points determining the 'hot point' when a reader can be converted to a paying version. Similarly, *Poool* allows for requiring payment only for certain sections of the online newspaper, to charge for articles for regular readers or to remove the paywall for occasional readers. It is based on an in-depth and incremental knowledge of each reader. The wider application of data and AI tools will enable content creators and producers to understand the impact of specific content on the audience and thus make better financing, production and dissemination decisions.

Automated Translation. AI can also be handy for automated translation for international broadcasters such as *Deutsche Welle* or *Arte*, allowing them to engage with wider audiences and increase their reach - although human polishing is still needed for professional quality translations.

All these innovative techniques and methods are changing (and promise to change further) news media and journalism. As such, they fundamentally concur to influence the algorithmic public opinion.

Concerns Related to the Algorithmic Public Opinion

There are several concerns related to the rise of algorithms in the media landscape. This chapter provides a summary of the major concerns. To begin with, market concentration in digital environments have led to the concerns that only a couple of algorithms have unprecedented and unaccountable influence on the information environment and news consumption of millions if not billions of people worldwide (algorithmic gatekeeping). Secondly, the black-box nature of algorithms raise concerns over the challenges of making these systems transparent, explainable and ultimately accountable (algorithmic opacity). Strictly intertwined, there is the inevitable risk that algorithmic systems are biased and lead to a variety of discriminations (algorithmic bias). Furthermore, there is a long-standing concern that algorithms – and more generally online social networks – can decrease the diversity of information to which individuals are exposed to and that eventually consume (so-called filter bubbles and echo chambers). The highly personalized data-driven and algorithmically-mediated information flows also trigger concerns over the risks of amplifying 'fake news' (disinformation) and creating addictive patterns of consumption (algorithmic manipulation). Finally, there are a number of equally significant concerns that are here briefly mentioned such as favoring a superficial news consumption, distraction, hate, and narcissism.

Algorithmic Gatekeeping

Nowadays few social media platforms – in particular Facebook, Youtube and Twitter – dominate most of the informational online media traffic. It is nowadays agreed in the literature about competition in social networks that an equilibrium can sustain only a small number of such intermediaries and a concentrated market structure is thus expected. Economies of scale in the production of news may indeed lead to monopolies. This is mainly because of network effects which occur when the value of a platform to any user increases exponentially with the number of already present users. As a matter of fact, innovation in this context has been pretty limited and the history of alternative social media is a history of failures. This is an inevitable outcome as long as visibility and scalability depend on economic and political resources (**Fuchs and Marisol, 2015**). The result is that few companies are de facto an oligopoly and, through their algorithms, they developed and preserve an immense 'opinion power', defined as the ability of the media to influence processes of individual and public opinion formation (**Helberger, 2020**). In addition to this, algorithms for content moderation also allow 'algorithmic censorship' (**Cobbe, 2020**). This power should come with certain responsibilities and legal liabilities (as it will be discussed in the section 'Editorial Obligations and Intermediary Liability'). How these companies design their platforms, how they allow content to flow, and how they agree to exchange information with competing platforms have clearly direct implications for both human rights

and innovation. The intermediation of content is related to several conditions of democracy; how people receive news, the articulation of relationships and associations; access to knowledge, and spaces for deliberation about issues of public concern. Nevertheless, platforms have always firmly argued that they are technology companies rather than media companies, thereby avoiding media regulation and editorial responsibilities (**Napoli and Caplan, 2017**). Algorithms, however, are not their only source of power; social media policies, technical design choices and business models also serve as a form of 'privatized governance' directly enacting rights and regulating the flow of information online and, in doing so, promote or constrain civil liberties (**DeNardis and Hackl, 2015**)

Distortive effects must be assessed in the context of the economically oriented algorithms. Especially with regard to social media, the main imperative behind personalization algorithms is indeed to "increase the number of regular users and the time that they spend on the platforms—to increase engagement" (**Napoli, 2019, p. 36**) because for platforms, it is primarily important to keep as many users for as long as possible to—corresponding to their business model—sell ads. There are indeed at least two main differences between traditional journalistic curation and algorithmic curation: on the one hand, relevant editorial news values (such as controversy, negativity, and elite people) interact with each other. A single factor is never decisive—unlike on Facebook, where only popularity with the users and their personal network mainly determines the content of the news feed. On the other hand, the basic direction is fundamentally different: while the news values that have traditionally guided journalistic gatekeeping emphasize social significance, the news values of intermediaries like Facebook focalize personal significance and are thereby primarily audience-oriented. To fully understand these dynamics, however, it is necessary a certain level of algorithmic accountability. This leads to the issue of algorithmic opacity.

(see → Box 2 ***The Case of Facebook's Algorithm***)

Algorithmic Opacity (or 'The Black-box Problem')

A primary source of concern is represented by the opacity of algorithmic systems. The understanding of algorithms in the academic world is weak due to two major factors: the impermanence of Internet technologies and the black-boxed nature of most influential algorithms (see **Bodo et al., 2017**). The first means that by nature the Internet is transient, rapidly changing at a rate that usually outpaces the research process. Algorithms are highly mutable. Google, for example, changed its algorithm 4887 times in 2020.[2] Secondly, the black-boxed nature of algorithms occurs not only to protect trade secrets but also prevent malicious hacking and gaming of the system (**Pasquale, 2015**). Of course, these features make not only research but also the policy focus on algorithms very challenging.

Despite their relevance, algorithms remain mysterious, invisible and mostly unknown. Not only they are protected by trade secrets but they are often inscrutable even to their creators. This is the problem

2 https://www.italian.tech/2021/05/23/news/le_ricerche_di_google_e_il_senso_della_vita-302295003/?ref=RHTP-BG-I302503236-P6-S3-T1

of *interpretability*[3] (**Albanie et al., 2017**). Similarly, it is problematic to unveil the reasons why algorithm made a specific choice. This is the problem of *explainability*[4].These create structural problems for *algorithmic transparency* and *accountability.* Furthermore, personalization algorithms adapt to one's behavior and expressed preferences and, therefore, they could even be considered unique to individuals. All in all, algorithms are complex socio-technical assemblages that are pervasive but still invisible, inscrutable, highly mutable and often adapted to an individual's profile or a specific context or dataset: these features make them an extremely difficult object to grasp, research and, eventually, to govern. Researchers and civil society have attempted to intepret the values inscribed into these algorithms and their consequences, especially through Applications Programming Interfaces (APIs). This allowed a number of relevant studies. Yet, in the aftermath of the Cambridge Analytica scandal, social media platform providers such as Facebook and Twitter have severely restricted access to platform data via their APIs, what **Bruns (2019)** ironically called 'APIcalypse'. This has had a particularly critical effect on the ability of social media researchers to investigate phenomena such as hate speech, trolling, and disinformation campaigns, and to eventually hold the platforms to account for the role that their affordances and policies might play in facilitating such dysfunction. Alternative data access frameworks, such as Facebook's partnership with the controversial Social Science One initiative, represent an insufficient replacement for fully functional APIs. Many researchers have argued the need for a return to web scraping[5], and a growing number of practical tools for scraping data from Facebook, Twitter, and other platforms are now becoming available. This approach, however, is inherently problematic. In addition to its dubious legal status, it is ethically questionable. This institutionalized algorithmic opacity is eventually one of the paramount concerns: if algorithms cannot be accountable, how can we trust them?

Algorithmic Biases and Discrimination

Algorithmic bias describes systematic and repeatable errors in a computer system that create unfair outcomes. In the context of automated content moderation there is an obvious risk of amplifying biases present in the real world. It is well documented that datasets are susceptible to both intended and unintended biases. How specific concepts are represented in images, videos, and audio may be prone to biases on the basis of race, gender, culture, ability, and more. Automated content moderation perform also poorly when tasked with decisions requiring an understanding of context and when analyzing new, previously unseen types of multimedia. Furthermore, the lack of diversity in engineering and design teams is an additional contributing factor to algorithmic bias. All this can eventually lead to discrimination against users. For example, models trained on several of the most widely used hate-speech datasets are up to twice as likely to label tweets by self-identified African Americans as toxic.

Content recommendation is also not immune to algorithmic biases.

3 Interpretability is the degree to which a human can understand the cause of a decision or consistently predict the model's result. In other words, how accurate a machine learning model can associate a cause to an effect. Interpretability should not be confused with the concept of explainability. The former is about being able to discern the mechanics without necessarily knowing why. The latter is being able to quite literally explain what is happening.

4 Explainability is the extent to which the internal mechanics of an algorithmic system can be explained in human terms.

5 Web scraping refers to bots which crawl web pages simulating human Web surfing in order to collect specified bits of information from different websites.

Even if the majority of users believes that algorithms are selecting content neutrally and informing impartially (**Gillespie, 2014**), the design of algorithms is inevitably affected by choices made by designers, i.e., which factors to include in the algorithm, and how to weigh them. They indeed use supplied criteria to determine what is "relevant" to their audiences and worth recommending, though these biases are not generally recognized.

The most known ones are the following:

Popularity bias

Information intermediaries often include popularity metrics in their ranking algorithm. A search algorithm, for instance, can give more weight to information coming from popular websites, to support majority interests and values. As a result, users may have troubles finding the less popular and smaller sites, the so-called long tail of content.

Third-party influence/manipulation

Because the information filtering are automated, they might be manipulated by activities from third parties. This could occur in several ways, from clickbait and Search Engine Optimization (SEO) techniques to 'bots' that game social media's metrics in order to further the spread of potentially problematic content.

Product/service prioritization

Studies showed that Google and Bing search engines both reference their own content in its first results position when no other engine does **(Bozdag, 2013)**. Facebook was also criticized for favoring the products of its partners. In the last decade, the EU received complaints that claimed how their traffic drop after Google began promoting its own services above conventional search results.

Novelty bias

In Google search engine, the number of years a domain name is registered has an impact on search ranking; domain names that exist for a period of time are preferred over newly registered ones. In Facebook, the longer a status update has been out there, the less weight it carries. A news item is prioritized over an old item. This might, for instance, lead companies to post updates when their audience is most likely to be online and using Facebook.

Other biases

The algorithm can also prioritize certain types of information over others. For instance, the engagement maximization business model of mainstream social media tends to prioritize ephemeral content to durable one, short videos (i.e., snippets) to long ones and casual images

(i.e., snapshots) to written text as they better lock-in users in the so-called "walled-gardens".

Media Diversity: Filter Bubbles and Echo Chambers

(see → Box 3 *The Challenges of Media Diversity*)

The impact of algorithms on the diversity of information has mostly been discussed negatively, assuming that such systems limit the breadth of viewpoints and topics. The major risk is the creation of "informational bubbles": filter bubbles (**Pariser, 2011**) and echo chambers (**Sunstein, 2017**), two sides of the same token. The first is a kind of cultural and ideological bubble in which an individual continues to see, listen and read what reinforces its opinions and interests. The latter is a group situation where established information, ideas, and beliefs are uncritically spread and amplified, while dissenting views are ignored. The crucial difference is that the former may not depend on the user's autonomy and awareness – therefore it is mainly caused by technological affordances – while the latter pre-exists the digital age and thus it is primarily driven by social relations.

From an individual perspective, content recommendation might reduce opportunities to self-determine and could negatively affect information finding by reducing the exposure to alternative points of view in the "marketplace of ideas" (**Pariser, 2011; Sunstein, 2017**) and, more generally, to serendipitous encounters (**Reviglio, 2019**). More generally, the main consequence to provide a 'too familiar world' is that our online life would eventually shift from an intersubjective to a subjective one (**Keymolen, 2016**). The consequences may be various: from the limitation of personal creativity to a reduction in the ability to build productive social capital (e.g. weak ties).

From a collective perspective, content recommendation can especially weaken media pluralism. As such, the audience would become increasingly politically fragmented and polarized and people – especially the less skilled and literate – more vulnerable to censorship and propaganda or, better, to "self-censorship" and "self-propaganda". This, in turn, would contribute to spread misinformation (**Vicario et al., 2016**) and erode interpersonal trust (**Keymolen, 2016**). It should be noticed, however, that in nations where political power is divided among several parties – and not only two like in United States – political polarization is more difficult to measure, and it is unclear whether it would even be possible.

There is limited evidence of the existence of these phenomena. Critics argue that these concerns mostly represent moral panics, and that personalization could instead foster the cultivation of "expert citizens" with stronger group identities. Also, they are poorly defined and, in fact, are used more as generalizing (thus limiting) metaphors (**Bruns, 2019**). To this date, research has been often contradictory, ambiguous and, ultimately, unconvincing (**Zuiderveen Borgesius et al., 2016; Bodo et al., 2017; Tucker et al., 2018**).

A number of challenges and concerns, however, persist. Firstly, these

phenomena are very hard to prove. Most research is often inconclusive because it is generally survey-based, or is correlational or based on small or unsatisfactory samples. In light of the fast-changing media landscape, many studies become rapidly outdated. Nevertheless, the risks of growing the "digital divide" and "cultural divide" or "epistemic inequality" remain. Certain privileged group of users, that have higher (digital) literacy, are able to manage more fruitfully personalization systems and online information consumption. Instead, a larger group of users would risk to be exposed only to a minimum, qualitatively inferior, range of information. Also, the wealthier the social networks, the more the benefits of personalization, and vice versa.

While insights on the main causes and risks of content recommendation have been currently understood (see **Tucker et al., 2018**), we still lack evidence with regard to the extent of their consequences. The issue is well beyond simplistic accounts that blame the algorithms. Information filtering processes take place not only at the technological level (both algorithms and affordances) but also on the individual (e.g., selective exposure) and the social (e.g., sharing practices). And given the vast heterogeneity of users, causes and effects of content recommendation vary widely.

Algorithmic Manipulation

The manipulation of our perception of the world is taking place on previously unimaginable scales of time, space and intentionality. There is indeed plentiful of information wars among several actors — state or non-state political actors, for-profit actors, media, citizens, individually or groups (see → **Malicous Actors**) — to the detriment of citizens and society at large. The risk of harm includes threats to democratic processes, including electoral integrity, and to democratic values that shape public policies in a variety of sectors, such as health, science, finance and more. The consequent misinformation can lead to public preferences different to if they were accurately informed, which can have negative policy implications. The same is true with public opinion more generally, where policy outputs feed back on public inputs into the policy-making process (**Dalton and Klingemann, 2007**).

Bots, "fake-news"[6] and political micro-targeting are the primary weapons of propaganda and, possibly, manipulation. Social Media in particular, and the Internet more generally, face accusations of deteriorating civil debate to the point that facts and truth are now fertile ground for dispute and subjectivity, while trust with experts and authorities have decreased, leading to a "post-truth era". More generally, today's Internet —especially social media (**Deibert, 2019**)— can be manipulative by design threatening individual autonomy (**Gal, 2017**; **Zarsky, 2019**). Human behavior can indeed be manipulated by priming and conditioning, using rewards and punishments. Eventually, the techniques employed can affect individuals' self-control, self-esteem and even self-determination. and can stimulate users in a powerfully subconscious and hormonal way. Facebook's large-scale emotional

6 Fake-news, intended as "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al., 2018), is considered by most academics as a vague buzzword - not to be confused with overlapping and better defined concepts like: misinformation – "which is information that is false, but not created with the intention of causing harm" – disinformation – "which is information that is false and deliberately created to harm a person, social group, organization or country" – and malinformation – "which is information that is based on reality, used to inflict harm on a person, organization or country" (Wardle and Derakhshan, p.20, 2017).

contagion experiment exemplifies this point (**Kramer et al., 2014**), showing how mainstream social media can affect emotions and exploit vulnerabilities in human psychology.

Since the Cambridge Analytica scandals implicating manipulative and possibly illegal social media use in Brexit and Trump 2016 campaigning, challenges and more effective solutions are being discussed. Despite apparently increasing transparency along with the resultant efforts to reform, social media often promoted extreme, inaccurate and radical content—regardless of what malicious actors may do to seed it. For example, one of the most radicalizing effect is the "rabbit hole effect" on Youtube, when algorithms capture a user in a spiral of ever more extreme (more often conspiratorial) content (**Yesilada & Lewandowsky, 2021**). Manipulation may occur also with political microtargeting, especially with 'dark ads' which are not recognizable as ads at all (**Borgesius et al., 2018**). One of the actual risks is that a more or less certain percentage of 'persuadables' – people known to be particularly vulnerable to targeted messages – can shift elections. In the context of search engines, this is called Search Engine Manipulation Effect (SEME) and can shift voters' preferences by 20% or more (**Epstein and Robertson, 2015**).

Then, there is the ability to create compulsion loops which are found more broadly in a wide range of social media (**Deibert, 2019**). These can be triggered through techniques such as variable ratio reinforcement[7] or A/B testing[8]. Such techniques become more effective thanks to affective computing (or 'emotional AI'), captology—the study of computers as persuasive technologies (**Fogg et al., 2002**)—and the emergence of psychographic techniques, along with diverse types of data such as location-based tracking, real-time data, or keyboard usage. Consider that suffice only a dozen of Facebook Likes to reveal useful and highly accurate correlations, such as predicting personality type, even better than one's parents prediction (**Youyu et al., 2015**). In addition, algorithms can even autonomously explore manipulative strategies that can be detrimental to users (**Albanie et al., 2017**).

At the same time, even design choices can be used to implement deceptive functionalities that are not in the user's best interest (so-called dark patterns, see → 'A Ban to Dark Patterns?') (**Gray et al., 2018**). Design facets can also intentionally trigger dopamine rushes or other emotional highs, stimulate popularity contests or implicit social obligations (**Kidron et al., 2018**). Such ability to nudge is defined by Yeung (**2018**) as a technique of "hyper-nudging" which dynamically configures the user's informational choice context in ways intentionally designed to influence her decisions. As such, hyper-nudging concerns all of the design process, not only algorithmic decision-making. These kinds of nudging techniques are already concerning in the case of negative effects on children's wellbeing, including increased risk of suicide and depression, conflicts with parents and adverse effects on cerebral and social development (**Kidron et al., 2018**).

7 Variable ratio reinforcement is a technique in which rewards are delivered unpredictably. This unpredictability affects the brain's dopamine pathways in ways that magnify rewards. It occurs when, after X number of actions, a certain reward is achieved (like in slot machines). In a personalized news feed, for example, among predictably uninteresting content is recommended a predictably 'serendipitous' content

8 In an A/B testing, the experimenter sets up two experiences: 'A,' the control, is usually the current system and considered the "champion," and 'B,' the treatment, is a modification that attempts to improve something—the 'challenger.' Users are randomly assigned to the experiences, and key metrics are computed and compared until it is found the versions that best exploit individuals' vulnerabilities.

There are several other intertwined concerns of the algorithmic public opinion as it is generally designed nowadays. Above all, the business model of mainstream social media is driven by the desire to capture the user's attention as much as possible. This is often referred to as the "attention economy", an economy in which user's attention is a commodity to exploit through ever more sophisticated techniques, including algorithmic ones. These certainly influence individual behaviors, information consumption and, therefore, the public opinion in many subtle and often contradictory and counterintuitive ways.

Superficiality and Distraction

Not only it is somehow deterministically assumed that information overload, multitasking and the disintermediation of information online are worsening our capacity of problem-solving and critical analysis (**Carr, 2011**), but algorithms optimized for maximizing engagement actually favor endless scrolling of information stimulating superficial news consumption. One increasingly common strategy employed by the younger generations to deal with this new environment, for example, is "news grazing", which occurs when users do not purposefully devote time to consuming news, maintaining a passive eye toward information, or keeping news sessions shorter. It is a form of news consumption motivated by the need for a wide breadth of information, rather than depth. It could be an adaptation to information overload. Such engagement might well be superficial if news consumers stop at just reading news headlines and short summaries. Algorithms can be designed to exploit this vulnerability and prioritize ephemeral content to durable one, as these are more engaging overall.

Narcissism and Online Participation

The quantification and maximization of social interactions contribute to create a "culture of performance" (**Castro, 2016**) that seems to be negatively correlated with well-being (**Verduyn et al., 2017**). Platforms like Facebook and Instagram do not manifest much as a public space but falls within the frame of the private and the exposure of the self. Among various consequences, narcissism thrives could lead to a de-politicization of society (**Byung-Chul Han, 2016**). Either way, it could lead to showing off political activism. A paradigmatic example is "slacktivism", which refers to the act of showing support for a political or social cause with very limited involvement required nor concrete effort. The main purpose is boosting the egos of participants in the movement. Popularity is the omnipresent tacit dream. In this rewarding-system machine, many - if not most - of the users show some compulsive behaviors. For example, many experience the fear of missing out from updates (what is called FOMO) so they would feel anxious without scrolling, login or internet connection. And despite such compelling engagement, the resulting environment is eventually toxic and exclusive, one that even suppresses the participation of most users. The majority of social media users are indeed lurkers, mere watchers of the online world, often falling into a "spiral of silence" in which they do not publicly express their opinions (**Noel-Neumann, 1984**). Platforms tend to

favor who plays their rules, and nudge users to play them.

Hate and Divisions

Incendiary and polarized content as well as hate speech are arguably widespread in social media. The consequences of this environment can be complex and various; from the dissemination of hate and incivilty to cultivating a subsectibility to indignation, from stimulating "on-demand politics" (where politicians deal with constant coverage and people's demand) to "identity politics" (when political agendas are based upon specific identities), from creating "culture wars" (cultural conflicts between social groups and the struggle for dominance of their values, beliefs, and practices) to ultimately leaving people distracted, divided and demoralised. If not "fixed", social media in the next years may increasingly facilitate identity-based violence. The risk, in fact, is that identity-based beliefs tend to eclipse truth-seeking because of the overriding need to belong and feel morally superior.

Main Actors in the Algorithmic Public Opinion

Before developing policy solutions focused on the prevention of negative consequences of algorithms it is paramount to make these systems transparent. But to make transparency unfold in algorithmic systems it should take into account not just code and data but an assemblage of human and non-human actors (**Ananny and Crawford, 2018**). This means to account for all the relevant actors in the algorithmic public opinion, most importantly users, policy-makers, governments, mainstream platforms and non-mainstream platforms, programmers and designers, moderators, influencers, journalists and news media, academics, malicious actors etc.. Of course, to maintain this policy review short, I present here only the main actors: policy-makers, platforms, users, news media, academics and malicious actors.

Policy-makers

To regulate or not to regulate? This is likely the main dilemma of policy-makers. The rapid development of the media landscape has notably left regulators to have to consistently catch-up with new innovations and inevitable trade-offs and conflicting human rights. This reality often leads to the "collingridge dilemma"; this argues that it is relatively easy to intervene and change the characteristics of a technology early in its life cycle. At this point, however, it is difficult to predict its consequences and regulators need convincing evidences before acting. Later, when the consequences become more visible and evidences are collected, it may be much more difficult to intervene. This is particularly true if we are dealing with the novelty and unprecedented scale of most algorithms in this context.

When a consensus to intervene is found, policy-makers are also faced with two interrelated questions: who should regulate and how these should be regulated. In general, there are three main approaches: traditional regulation, self-regulation and co-regulation who set and enforce different regulatory goals, standards and justifications (**Hirsch, 2010: Finck, 2018**). These reveal different approaches to regulation, yet they operate on a spectrum. Also, all of them are already set in place to some extent. On the one hand, platforms are already self-regulating entities; they determine the terms and conditions of their intermediary function and define online and offline standards of behaviour. On the other hand, there is an information asymmetry between platforms and policy-makers that naturally requires some forms of co-regulation which usually takes the form of voluntary codes of conduct and negotiated self-regulatory agreements. The state role, in this case, is relatively limited to more of an informal oversight and steering role.

Fundamentally, policy-makers lack a shared vocabulary or frameworks for approaching the incoming challenges, and new models for digital governance are likely needed to be developed. The algorithmic public opinion is in fact particularly challenging to policy-makers as it intersects with a transformation of media itself, due to media convergence, media globalization and the rise of global digital platforms. Traditional policies for broadcasting, telecommunications, and media are more often inadequate. There is a risk that "regulatory transference" – the application of existing law and regulation to new business models and market conditions – does not address the right problems, and has unintended consequences. Legal and regulatory frameworks designed for one set of market and technological circumstances may be ineffective or inappropriate in others. The complexities of contemporary digital systems and networks, algorithmic content filtering, data ecosystems, social media, and cross-platform activities necessitate different methods to address the issues and challenges they pose. Domestic policies can address some issues, but global policy is progressively more significant to address communication challenges and classic regulatory concerns such as compliance, enforcement, and efficacy.

In the past years policy-makers have undertaken several different initiatives to regulate social media platforms and their algorithms. In particular, this has included horizontal instruments, such as competition and data protection law, which are not specifically tailored to social media and their algorithms, but may still have some spillover benefits for certain purposes. Antitrust and consumer protection law, in fact, already place certain limits on platforms, for example limiting their ability to prioritize their own services without discriminating other actors or even forbidding covert advertising, which means that platforms have a duty to disclose whether content is being sponsored.

To date, (supra)national governments have tried some forms of co-regulation, such as the EU Code of Practice on Disinformation, in which the major tech companies – Facebook, Google and Twitter – pledged to work more actively to lessen the spread of disinformation and hate

speech online. These, however, are usually non-binding and the rules could be interpreted rather loosely by companies. Over the years, many critics in academia and civil society have argued that co-regulation and self-regulation attempts do not provide sufficient incentives to act in the public interest. It is nowadays widespread these need to be replaced by the law (**Floridi, 2021b**).

Policy-making also presents a number of structural limitations that limits its decisive role. To begin, there are legitimate but concerning lobbying activities to take into account. Even the often praised European General Data Protection Regulation (GDPR) is one of the most lobbied pieces of EU legislation to date (**Edwards and Veale, 2017**). Lobbying is very strong even for new European legislative drafts such as the Digital Services Act and the Digital Markets Act. The resources these companies can employ to negatively affect innovative policies are huge. This brings concrete risks that need further analysis to, eventually, cultivate a more effective public oversight. Similarly, another phenomenon is the one of revolving doors which refers to the movement of personnel between roles as legislators and regulators on one hand, and members of the industries affected by the legislation and regulation on the other. For example, The Google Transparency Project has so far identified 258 instances of revolving door activity (involving 251 individuals) between Google or related firms, and the federal government of US, national political campaigns and Congress during the Obama's presidency. Bank et al. **(2021)** argue that these should be simply blocked. Finally, there are political pressures that might worsen deliberation. Notably, what Ananny and Gillespie **(2016)** call "governance by public shocks". They argue that social media regulation is often driven by "public shocks", these are shocks that "sometimes give rise to a cycle of public indignation and regulatory pushback that produces critical - but often unsatisfying and insufficient - exceptions made by the platform (p.3)." Such reactive approach from politics works against proactive, long-term policy solutions.

Mainstream Platforms

Today's social media landscape is characterized by an oligopolistic market in which a small group of platforms — above all Alphabet (Google) and Meta (Facebook)— act as the ultimate gatekeepers for billions of users worldwide. The European Digital Service Act refer to these as "very large online platforms", the ones with more than 45 million monthly active users, providing special rules. When we talk about platforms we in fact refer to the these companies which provide other than Google search and Facebook other fundamental online services such as Google maps, Google news, Youtube, Whatsapp, Instagram and many others.

These companies have created a global architecture of data capture and analysis produces rewards and punishments aimed at modifying and commoditizing behaviour for profit. Basically, they harvest large amounts of data (i.e. big data) to identify patterns of consumer

behaviours, preferences, tendencies, etc. This business allows the existence of "free" services. In order to access any social service platform, however, users have to give up the social data they generate and accept to be surveilled to help algoritms predict their desires through personalized ads. It is indeed the ad-industry the main source of revenue of these companies. In 2017, advertising constituted 87 percent of Google's total revenue and 98 percent of Facebook's total revenue. This is what is often referred as 'surveillance capitalism' (**Zuboff, 2015**) in which the capitalistic logic of accumulation produces "hyperscale assemblages of objective and subjective data about individuals and their habitats for the purposes of knowing, controlling, and modifying behavior to produce new varieties of commodification, monetization, and control" (**Ibid., p.85**). There is indeed an economic and political antagonism between users' interest in data protection and corporate tax accountability on the one side, and corporations' interest in user data's transparency/commodification and corporate secrecy on the other side.

Mainstream platforms provide a personalized experience, but very little control over information filtering processes. This control asymmetry perpetuates an epistemic imbalance in which platforms know people more than people know themselves. Many of the practices involved in surveillance capitalism are challenging social norms associated with privacy and, thus, are contested as violations of rights and laws. Consequentially, these corporations have learned to "obscure their operations, choosing to invade undefended individual and social territory until opposition is encountered, at which point they can use their substantial resources to defend at low cost what had already been taken" (**Ibid., p.85**). On their side, mainstream platforms defend themselves through a mix of trade secrets, economic claims, promises of self-regulation, and technological solutionism, forestalling real public oversight. They essentially respond to regulatory threats by claiming that:

1. Their systems use proprietary knowledge that they cannot publicly disclose;
2. Their business models require large-scale data harvesting;
3. People are unwilling to pay for services that are currently ensured by people's data;
4. Encryption technologies, transparency commitments, and controlled data disclosure obviate the need for public oversight.

It is important to understand why and how such model became dominant. According to Zuboff **(2015)** there are a variety of reasons. Firstly, this system was constructed at high velocity and designed to be undetectable. Structural asymmetries of knowledge and rights, in fact, made it impossible for people to learn about the above practices. There was no historic precedent, so there were few defensive barriers for protection. Leading tech companies have been over-estimated, in a way respected and treated as 'emissaries of the future'. On the other hand, individuals quickly have become to depend upon the new information and communication tools as necessary resources,

requirements – at times even preconditions – for social participation. As Zuboff argues: "the rapid build-up of institutionalized facts (...) produced an overwhelming sense of inevitability (p.85)." And as the philosopher Mireille Hildebrandt (**2015**) rightly claims: 'the temptation to accept that things are the way they are because technology is the way it is, has a strong hold on public imagination [...] once a technology has consolidated it acquires a tenacity that is not easily disrupted.' (p. 174).

This model should lead – and to some extent is indeed leading - policy-makers to consider more radical policy solutions to tame the unprecedented, institutionalized, platforms power. Considering that these companies represent key nodes in economic, social, and informational flows it has been argued that they could even be treated as public utilities (**Rahaman, 2018**) while the data generated by users as a public resource (**Napoli, 2019**). Of course, it is not that simple; there are a number of fundamental and complex systemic issues related to the power of these companies that will be introduced and further discussed in the next Chapter.

Users

Users are the ultimate gatekeeper. This is why it is mostly stressed that for any concern related to the algorithmic public opinion what is most needed is literacy; not only media and algorithmic literacy but also digital skills, including an awareness of the challenges of digital societies. Users can also become active players, helping platforms and authorities in many ways: from flagging content to self-organized troops against disinformation online and, more generally, demanding more transparency and autonomy of choice.

The algorithmic system which is arguably the one in which users can act more substantially upon is certainly content recommendation. In this context, it has often been stressed the primary role and therefore responsibility of users as a prominent argument to argue against strong regulation. Many of the problems with media pluralism are indeed mostly user-driven (**Helberger et al., 2018**). It is therefore necessary to understand users, their behavior and their perceptions. These, in fact, deeply influence their trust on platforms and, in turn, how they use them. Generally, users across the world seem to embrace personalization algorithms as they enjoy the services they offer on a daily basis. (**Newman et al., 2016**). As said, for the majority of users news algorithms are not doing anything wrong: they select neutrally and inform or recommend impartially (**Gillespie, 2014**). As such, most people trust more themselves in delegating to algorithms than they trust journalists. Users indeed guess the functioning of algorithms through what are called "folk theories", developing a more or less accurate "algorithmic imagination" (**Bucher, 2017**). However, most users still miss the fact that there is no objectivity in the realm of algorithms. Personalization algorithms embodies specific forms of power and authority. Still, most users have no reservations in principle against news being distributed through AI-driven tools (**Thurman et al., 2019**).

Yet, they increasingly tend to be worried that "more personalized news may mean that they miss out on important information and challenging viewpoints" (**Newman et al., 2016**, p.113). Research indicates that they expect to play a more active role in the interaction with AI-driven tools such as personalization systems (**Helberger et al., 2019**). Also, there are concerns about being wrongly profiled, as well as privacy concerns.

In short, the majority of users trust algorithms but less and less and they enjoy every day their fundamental services more and more. Though they can have some reservations about algorithms, many have even more reservations about journalism and editorial selection. Indeed, journalism is in crisis and more often survives with click-baiting. As Internet critic and scholar **Evgeny Morozov (2017)** puts it "what it has gained in profitability, it seems to have lost in credibility." Trust is fundamentally at stake in these processes. Nevertheless, perceptions, expectations and behaviors over (news) content recommendation remain rather nuanced, ambivalent and at times even inconsistent.

There are two main explanations for the widespread default trust and lack of individual and institutional responsiveness over platforms and, in particular, their opaque and limitedly controllable personalization systems. On the one hand, as said above, one compelling explanation is that nothing similar to ICT development ever happened in the past, so there were few institutional and individual defensive barriers (**Zuboff, 2015**). Societies quickly came to depend on these new information and communication tools as necessary resources, and at times even as preconditions for social participation. On the other hand, users' behavior has to be analyzed in light of behavioral economics and psychology (**Acquisti et al., 2015**). Drawing from the work of Kahneman **(2011)**, Thaler and Suntein **(2019)** describe how our bounded rationality affects how we assess the likelihood of future events and how our individual biases and vulnerabilities could be exploited. This particularly explains the so-called "privacy paradox" – where people claim to value privacy but they don't protect it proactively – yet it can be applied in many other circumstances such as individual control over personalization systems or vulnerability to dark patterns.

More generally, the use of AI-driven tools fundamentally alters the agency of users regarding the news individuals consume. AI-driven tools have introduced observable measurements of user interaction allowing detailed insights into audience preferences, impossible to obtain in non-digital media. Even if users are at the center of these processes, this increase in user agency is unidimensional as it is solely focused on observable engagement like clicks or time-spent. And as explained in the previous chapter, users do not know what information they are automatically excluded from. The challenges outlined above lead to a fundamental open questions: can users be sufficiently informed to be fully responsible in such a fast-changing complex media environment? To what extent and in which forms they ought to be supported?

News Media

News Media and its adaptation to the digital era is fundamental to understand the actual algorithmic public opinion. It is indeed well-known that news media and journalism are in a serious economic and trust's crisis. Over the last two decades, as people increasingly began to consume news through smartphones and other electronic devices, a shift in the consumption of print media channels has occurred. News from a variety of online sources, like blogs and other social media, resulted in a wider choice of official and unofficial sources (e.g. citizen journalism), rather than only traditional media organizations. One of the consequence of this is that most people expect online news to be free and do not buy newspaper anymore. News media – and especially newspapers – have therefore seen print revenues heavily decreasing, eventually reducing their staff and coverage. News organizations have been challenged to recover losses and find new business models for making journalism sustainable.

The industry is struggling and dealing with this crisis in different ways. One initial widespread reaction has been favoring quantity over quality and sensational titles (so-called click-bait). Through these practices, news media earned the online information industry the epithet of "Clickbait Media" (**Munger, 2020**). Another approach has been the one of paywall, a method of restricting access to content, with a purchase or a paid subscription. Beginning in the mid-2010s, newspapers in fact started implementing paywalls on their websites as a way to increase revenue, partly due to the use of ad-blockers. While this approach has its merits, especially in recovering lost revenues, critics argue that erecting paywall restricts equal access to the online public sphere.

One of the consequences of the current news media crisis has been a decline in trust. In recent decades, trust in news has indeed declined in many parts of the world (**Newman, 2020**). This decline in public confidence in the press is part of a broader skepticism that has developed about the trustworthiness of institutions more generally – state, science and interpersonal – leading to an overall concerning trust recession. This has led to discussions on how journalists define and enact their democratic role and how news organizations ought to give journalists the freedom and encouragement to engage in trust-building experiments to facilitate discussion, building community, and partnering with the public.

As prerviously argued, news media is also confronted with the innovations and challenges brought by algorithms. Notably, these have led to data-driven journalism (or data journalism), a journalistic process based on analyzing and filtering large data sets for the purpose of creating a news story. Data journalism reflects the increased role that numerical data is used in the production and distribution of information in the digital era as well as the increased interaction between journalists and several other fields such as design, computer science and statistics. Deciding what's news is indeed increasingly influenced by quantitative

audience measurement techniques. In particular, the constant audience measurement returns real-time and personalized data on practices consumption of the public. Data are therefore often seen as a primary avenue towards the sustainability of journalism in the digital era.

The result of this comprehensive datafication process applied to journalism manifests itself in at least four forms (cfr. D6):

1. Journalistic outcomes are increasingly a byproduct of algorithms in terms of positioning in the search engines and in the timelines of social platforms. Widespread practices of search engine optimization (SEO) makes journalism adapt to the algorithms' will.
2. The metrics of engagement with journalistic content are available to everyone (including readers): how much an article is shared, how many likes it receives, what is its circulation. This can have different, mixed implications: from deepening an issue that interest the audience to adapt their writing style based on users' comments. In any case, it implies a more audience-driven journalism.
3. Journalistic metrics are a product of a larger analytics processing system that monitors individual and aggregated behavior (i.e. Google Analytics, Charbeat, Newsbeat, Parse.ly). As Diakopoulos (2016) argues, this has led to an increase in directly related metrics to the growth of digital platforms in which the act of consuming news generates a different transmission of data.
4. Journalistic metrics as a product of behavioral processing of network users (such as NewsWhip, Crowdtangle, Ezyinsights). These softwares monitor what is happening on the Internet, keep track of social media signals, monitoring tweets, shares and comments. These new forms of journalism are not fully understood yet, but they significantly concur to the development of an algorithmic public opinion as here conceived.

Academia

Academia has a special role in the algorithmic public opinion. Being at the forefront of the understanding of the digital revolution, it helps to raise civic awareness, providing theoretical critiques, scientific evidences, engaging in public advocacy and do consulting to policy-makers and businesses. Its independence and effectiveness are, however, questionable.

Scholars have developed and have been involved in plenty of meaningful initiatives; building new tools. For example, *Twitter Capture* and *Analysis Toolkit* (TCAT), developed at the *Digital Methods Initiative* (DMI) at the University of Amsterdam; providing public oversight on various issues, from social bots to misinformation. Another example is the *Computational Propaganda Project* by the Oxford University which provides meaningful updated analysis and data for anyone interested; providing critical insights on new legislative proposals. Another paradigmatic initiative is the *Digital Services Act (DSA)* Observatory of the

Institute for Information Law of the University of Amsterdam, a project exclusively dedicated to the monitoring of the proposed new European law DSA; finally, scholars can be part of research and ethics boards, notably the philosopher at the Oxford University Luciano Floridi was initially part of the *Google AI Ethics Board* that was subsequently shut down.

Building on increasingly unprecedented large datasets, gathered through the *Application Programming Interfaces (APIs)* of mainstream platforms, social media research had proven a particularly fertile field: the scholarly community developed and shared increasingly sophisticated methods and tools for gathering, analysing, and visualising social media data. There is a vast potential for social sciences. Yet, as previously mentioned, in the aftermath of the Cambridge Analytica scandal, social media platform providers such as Facebook and Twitter have severely restricted access to platform data via their APIs, what **Bruns (2019)** ironically called "APIcalypse". This has had a particularly critical effect on the ability of social media researchers to investigate phenomena such as hate speech, trolling, and disinformation campaigns, and to eventually hold the platforms to account for the role that their affordances and policies might play in facilitating such dysfunction.

The question of API's and data access is paramount for researchers and it's becoming a matter of increasing contention. Over the past year, dominant platforms such as Facebook have repeatedly interfered with independent research projects, prompting calls for reform. In October 2020, before the US elections, Facebook tried to shut down an independent audit of their political advertising by NYU. Similarly, it has retaliated against data collection by the NGO *AlgorithmWatch,* sending them threats of legal action on the grounds that independent data collection violated the platform's Terms of Service. Given the considerably asymmetrical distribution of resources and control over platform affordances between providers and scholars, it is therefore difficult to see this as a race that can possibly be won by researchers, even if the legal, moral, and ethical justifications for entering into it can be found.

Scholars too, similarly to policy-makers, can be tempted by mainstream platforms in "revolving doors" or somehow hushing them. Over the years, the platforms have also hired a substantial number of university graduates – in instrumental fields such as computer science but also from more critical disciplines such as media, communication, and cultural studies – to bolster their workforce in key areas of operation; at times, they also collaborate with external research teams. And from time to time, platforms offer 'data grants' and similar competitive schemes that explicitly invited scholars to apply for funding and data access for particular research initiatives. The role of academia for the understanding, development and governance of the algorithmic public opinion is undoubtedly significant and multifaceted. Yet, the potential to collect meaningful evidences is seriously limited by data access. Furthermore, the essential role of informing stakeholders makes

scholars critical actors in this context.

Malicious Actors

The algorithmic public opinion has favored new opportunities for malicious actors; many kinds of profit-driven and especially politically interested actors are finding ways to exploit social media and algorithms, in particular to spread propaganda, a phenomenon often reffered to as 'the weaponization of social media'. Operations can be conducted by government agencies, politicians and parties, private companies, influencers and citizens. Foreign influence operations, disinformation, and state-sponsored trolling and harassment have indeed already undermined human rights and degraded the quality of political news in circulation.

Since 2016 the *Computational Propaganda Project* of the Oxford University has monitored the activity of "cyber troops", defined as government or political party actors tasked with manipulating public opinion online (**Bradshaw & Howard, 2017**). They have examined the formal organization of cyber troops around the world, and how these actors use computational propaganda for political purposes. As such, they have built an inventory of the evolving strategies, tools, and techniques of computational propaganda. The phenomenon is indeed concerning and steadily growing. Yet, many of these operations are almost certainly not even been publicly documented.

In their 2020 report, Bradshaw et al. have classified the valence and messaging strategies used by cyber troops into four categories: pro-government or pro-party propaganda, attacking the opposition or mounting smear campaigns, suppressing participation through trolling or harassment, drive division and polarize citizens. These strategies are conducted throughout a number of techniques:

1. The creation of disinformation or manipulated media. This is the most prominent type of communication strategy and includes creative so-called "fake news" websites, memes, images or videos, and other forms of deceptive content online even deepfakes.
2. Dark ads. These are data-driven strategies to profile and target specific segments of the population with political microtargeting.
3. Trolling, doxing or online harassment. In several countries have been found evidence of trolls being used to attack political opponents, activists, or journalists on social media.
4. Censor speech and expression through the mass-reporting of content or accounts. Posts by activists, political dissidents or journalists can be reported by a coordinated network of cyber troop accounts in order to game the automated systems social media companies use to flag, demote, or take down inappropriate content.

Cyber troops normally use accounts to spread computational

propaganda, for example: automated accounts (or political bots), which are often used to amplify certain narratives while drowning out other; human-curated accounts, which might use low levels of automation but also engage in conversations by posting comments or tweets, or by private messaging individuals via social media platforms; hacked, stolen, or impersonation accounts (including groups, pages, or channels), which are co-opted to spread computational propaganda, though they represent a small portion of the total. More generally, these cyber troops could also employ so-called click-farms in which are low-paid workers hired to click on ads but also to generate likes, followers and, more generally, engagement.

The algorithmic public opinion also entails cyber-espionage and 'information warfare' . In this context, for example, bots can also be used to extract data through web scraping which simulates human Web surfing in order to collect specified bits of information from different websites. These techniques can be used for "open-source intelligence" (or OSINT). OSINT is produced from publicly available information that is collected, exploited, and eventually disseminated to an appropriate audience for the purpose of addressing a specific intelligence requirement. This could, for example, result in sensitive lists that can be used as a form of political intelligence, as it was recently the case with a Chinese company who created a list of political important networks in Western countries simply crawling publicly available information (**Balding, 2020**). Furthermore, malicious actors can buy malwares in the black market (or even totally legit ones) to monitor people, notably activists, journalists, and political leaders. The recent Pegasus Project leak is a paradigmatic example (**Amnesty International, 2021**). Several governments have been accused to have exploited these malwares. These malicious tools further threaten freedom of speech. Anyone could monitor and threaten people who could or would speak up about issue of public interest.

Other serious concerns are data breaches and how these could be exploited by data brokers - companies that collect consumers' personal information and resell or share that information with others. In the hands of malicious actors, breached data can indeed become a tool for cyber espionage, political campaigns and, more generally, information warfare such as disinformation operations, especially during elections. This raises serious concerns on the effectiveness – or even enforceability – of data protection regulation at a global level. Data brokers in fact can proactively obfuscate the source of their data, making it difficult for any individual to retrace the paths through which their data were collected (**Reviglio, 2022**).

The Governance of the Algorithmic Public Opinion

The governance of the algorithmic public opinions entails the challenges and opportunities to re-design our societies. In fact, to re-think how information is produced and disseminated, how we mediate our relations online, how we get news and form opinions individually and collectively and, ultimately, how to increase the quality of public opinion and, therefore, of democratic deliberation. These represent fundamental issues for preserving democracies and human rights for the years to come. Indeed, the fast-paced media environment along with disruptive technological innovations (prominently artificial intelligence) and a number of epochal challenges (from climate change to increasing geopolitical tensions) require regulators, companies and civil societies to be able to continuously maintain cohesion and reach new agreements. This also requires finding agreements on new set of rules to develop a more innovative, transparent and inclusive Internet and social media environment. The following analysis, therefore, aims to help to shed light on the limitations and challenges of current regulations and policies. Firstly, the most critical debates for the governance the algorithmic public opinion are introduced. Secondly, relevant policy challenges and debates are mapped; on the one hand, systemic issues and the most prominent policy approaches for the governance of algorithms, antitrust, data protection and media liability are briefly introduced. On the other hand, more specific issues such as recommender systems, content moderation, social bots, political microtargeting and disinformation are discussed.

The Challenges of Governing the Algorithmic Public Opinion

For emerging digital technologies, legislation is a rather blunt tool; it always risks being too broad to be useful, or too specific to be future-proof. The Internet has also undermined the benchmarks that acentury of international cooperation helped to build; the principles of territoriality, universality of values and effectiveness in the cyberspacecollide with the fluidity of data, algorithms and information flows. Social media and their personalization algorithms are certainly not easy to regulate either; the question of regulating online information flows is part of a complex regulatory system involving many different legislative acts and normative approaches.

The question of how to effectively govern platforms and their algorithms is an open challenge. Together with the centrality of national authorities, it is widely shared the need to move away from homogenous top-down models towards decentralized, reflexive, effective, collaborative and cooperative frameworks that are ‘polycentric’ (see **Finck, 2018**). Regulatory conversations on social media platforms are indeed already polycentric in that they are transnational and multisectoral. Polycentricity is indeed inherent to new governance

models where no actor does make decisions unilaterally. Unlike traditional conceptions of law that rely on a unitary source of authority, new governance is based upon a dispersal and fragmentation of authority, and rests upon fluid systems of power sharing. In this context, it is essential, for example, to be able to combine digital regulation with ethics and governance. The regulation of the digital, the governance of the digital and the ethics of the digital (whether computer, AI, information or data ethics) are in fact different normative approaches, complementary, but not to be confused with each other (**Floridi, 2018**). Indeed, not every aspect of regulation is a matter of governance and not every aspect of governance is a matter of regulation. For example, governance may comprise guidelines and recommendations that overlap with, but are not identical to regulation, whereas ethics shapes regulation and governance through the relation of moral evaluation.

Platform Governance

Aside from theoretical approaches, there is also a more pragmatic need to coordinate regulatory efforts while innovating the governance of platforms. Importantly, "platform governance" - understood as the set of legal, political, and economic relationships structuring interactions between users, technology companies, governments, and other key stakeholders in the platform ecosystem (**Gorwa, 2019**) — is indeed rapidly moving away from an industry self-regulatory model and towards increased government intervention (**Helberger, Pierson, & Poell, 2018; Floridi, 2021b**).

As a matter of fact, platform governance does not depend exclusively on a single source of accountability or regulation, but rather on more complex and multistakeholder systems of governance. The complexity of governing the algorithmic public opinion, in fact, results also from the fact that a large number of different stakeholders are involved in the development, production, distribution, exploitation and marketing of social media content. The three main categories at stake are users, content providers/producers and distributors/platforms, which in turn are split into a number of different types of actors – much more diverse in the digital environment than in the analogue environment. The regulatory space has indeed increased dramatically due to the borderless nature of the digital world, as it has the technical expertise needed to create effective, appropriate and enforceable rules. For these reasons, the involvement of various stakeholders in regulatory approaches has become ever more important.

Multistakeholderism

One way to mitigate the challenges of the governance of the algorithmic public opinion is, therefore, to incorporate multistakeholder and co-regulatory elements. On the one hand, both governments and international organisations have often failed to produce adequate solutions to contemporary corporate transnational governance and policy issues. Traditional regulation - where governments seek to make corporations comply under threat of legal and financial penalties - it

is not always easy to pass, as industry lobbies heavily to protect its interests, and even once rules are in effect, ensuring compliance — especially when firms are headquartered in different jurisdictions — is no easy task. Thus, multistakeholderism becomes meaningful; different governance stakeholders have differing levels of regulatory capacity that they bring to the table: each type of actor has different competencies that are required at different phases of the regulatory process, from the initial agenda-setting and negotiations to the eventual implementation,monitoring, and enforcement of governance arrangements.[9]

"Cooperative Responsibility"

Helberger, Pierson and Poell (**2018**) argue that "the realization of public values in platform-based public activities cannot be adequately achieved by allocating responsibility to one central actor (as is currently common practice)" and therefore envisage a "dynamic interaction between platforms, users, and public institutions". Social media policy is in fact moving towards a 'cooperative responsibility', that is, the result of the dynamic interaction between platforms, users and public institutions. In such (re)distribution of responsibilities, they identify four key steps:

1. To collectively define the essential public values at play in particular contexts
2. Each stakeholder (platforms, governments, users, advertisers, and others) accepts that they have a role to play in the realization of these value
3. It is developed a multi-stakeholder process of public deliberation and exchange, in which agreement can be reached.
4. They translate the outcome of public deliberation and agreements into regulations, codes of conduct, terms of use and, last but not-least, technologies (e.g. 'by design').

Towards a New Governance

Co-regulatory, multistakeholder solutions that allow the dynamic constructive relationship between ethics, regulation and governance bear the potential to allow for more informed decision-making, easier enforcement, and continuous and accountable review and assessment. The experimental nature of this process allows for mutual learning and the identification of best practices as well as for a dynamic adaptation of the relevant rules over time. However, to believe that "a single effective and proportionate regulatory approach could be designed in such a way as to tackle every one of these matters is highly presumptuous and neglects the wide array of complex social factors underpinning the production, sharing and engagement of such content (**Nash, 2019, p.19**)." This review, in fact, has no ambition to provide any comprehensive governance approach but only to stimulate the policy debate by

9 It is also true that the governance of online content on platforms is a far less multistakeholder than the typical Internet Governance (IG) of internet protocols and standards, with far fewer formalised institutions and fora. To be clear, multistakeholderism is no panacea and civil society has been often marginalised in IG, eventually merely serving to legitimise the process for other, more powerful actors.

discussing promising normative approaches and policies proposals.

Systemic Issues

Digital media and the growing data economy has profoundly disrupted media markets, challenged the existing business, production and consumption of news, and presented fundamental challenges for pol cy-makers. The increasingly intertwined areas of algorithmic regulation, data protection, media and competition are large and complex yet essential to understand the full picture of the regulatory dynamics behind the algorithmic public opinion. In this Section, I identify and present fundamental policy challenges and developments that are particularly salient for the algorithmic public opinion. Of course, I do not indend to provide any comprehensive guide to the regulation of systemic issues. However, I aim to introduce the most significant debates and current proposals around systemic issues which are intended as complex regulatory issues that substantially (and sometimes indirectly) influence the conditions for a healthy and sustainable algorithmic public opinion ecosystem.

Algorithmic Regulation and Accountability

Algorithmic regulation – including issues related to the use of AI and big data – has risen up the public agenda in recent years, with a range of reports issued by international agencies, government departments, legislative commitees, think-tanks and academic bodies. Generally, attention has been given to a variety of issues such as raising awareness on the functioning and (unintended) consequences of algorithms (e.g. **Algo:aware, 2018**), the risks these entail for democracy (e.g. **AlgorithmWatch, 2020**) and for discrimination (e.g. **Zuiderveen Borgesius, 2018**) and, eventually, how to design algorithmic systems responsibly (e.g. **Alan Turing Institute, 2019**), use audit methods in social media (**Ada Lovelace Institute, 2021**) and assess the impact of algorithms (e.g. **AI Now Institute, 2018**). Similarly, an intertwined strain of analysis is concerned with AI more broadly; from ethics guidelines (for example, **HLEG AI, 2019**) and human rights perspectives (e.g. **EU Agency for Fundamental Rights, 2020**; **Fjeld et al., 2020**) to issues more relevant for the algorithmic public opinion such as how AI-driven tools in the media affect freedom of expression (**Helberger et al., 2019**).

Algorithmic regulation now features in so many different contexts that it must engage a variety of legal rules beyond the data protection sphere: administrative law, criminal law, intellectual property law, contract law, and competition law are obvious cases in point. There are also a large number of initiatives aimed at promoting responsible algorithmic decision-making, including working groups and committees, policy and technical tools, standardisation efforts and, finally, codes of conduct, ethical principles and frameworks (**Algo:aware, 2018**). Of course, it is unlikely that a single policy solution or approach will deal with all, or even most of those challenges discussed above. In order to address them, and to manage the tradeoffs

that arise, a layered variety of approaches are likely to be required. This section will therefore lay the ground for the subsequent analysis of specific issues, in particular recommender systems and political microtargeting, by introducing the main concepts and debates. Algorithmic regulation is an emergent field of governance. Civil society and industry have already begun to develop initiatives and design technical tools to address some the issues raised throughout this section. In particular, issues and questions around the benchmarks of accuracy, fairness, accuracy and accountability are fundamental for the development of policy toolboxes. During the 2010s the everincreasing omnipresence of AI systems has been accompanied by a growing importance attached to the ideals of fairness, accountability and transparency in relation to algorithmic decision making (so-called FAT); in a nutshell, fairness means the absence of systematic bias and disadvantage towards particular demographics or social groups; accountability is a tool that is supposed to contribute to fairness: it has to be accounted for the algorithmic activities and accepted the responsibility for the resulting (unfair) outcomes; transparency has to do with visibility and insight into the system. In addition, there are often discussed the principles of explainability – how AI outcomes can be understood by humans (also called Explainable AI or XAI, see e.g. **Adadi and Berrada, 2018**) and accountability – which refers to "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences. Thus an 'actor' (be they an individual, a group, or an organization) is required to explain their actions before a particular audience, the 'forum' (24, p. 447) (for a literature review see **Wieringa, 2020**). There are a number of techniques and approaches to verify the integrity of algorithmic systems.

Algorithmic Impact Assessments. Principles are employed in algorithmic impact assessments. An impact assessment can be defined as "the process of identifying the future consequences of current or proposed action."[10] By requiring an entity to conduct an internal inspection, impact assessments urge coders and designers to conduct a deeper form of analysis, carefully investigating plausible areas of bias, error, and uncertainty, as well as implementing the necessary steps to correct them. Of course, different jurisdictions have different impactassessment schemes in place, and each has its own specificities and objectives. Diakopoulos et al.[11] elaborated the Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. These are responsibility, explainability, accuracy, auditability, fairness. Generally, however, impact assessments may provide only limited transparency and allow only limited room for public review, for example for automated content moderation (see **Nahmias and Perel, 2021**).

Human Rights and Risk-based Approches. Many of the issues raised against principled-based governance can also be taken care of by applying a human rights lens to algorithmic systems. One of the earliest applications of the human rights framework to the topic of AI was the Toronto Declaration - Protecting the right to equality and

10 See the definition employed by the International Association for Impact Assessment https://www.iaia.org

11 https://www.fatml.org/resources/principles-for-accountable-algorithms.

nondiscrimination in machine learning systems (2018). Since then there has been an ever increasing amount of work in this area, with academics, civil society organisations and international bodies all publishing work on human rights and AI. Whereas voluntary ethics guidelines leave large scope for companies to interpret what different principles mean, the international human rights framework has established mechanisms for resolving such ambiguities, and enforcing compliance, even if that hasn't always been without issues. Another approach in this context is a risk-based one. These involve determining the scale or scope of risks related to a concrete situation and a recognised threat. This approach is useful in technical environments where companies have to evaluate their own operational risks.

Algorithmic Auditing. Auditing techniques are a key part of a regulatory inspection process to ensure accountability. Algorithmic auditing enables interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism. In the contex of social media platforms, auditing can be undertaken using a variety of techniques such as:

1. 'Code audits' (when auditors have direct access to the codebase of the underlying system);
2. 'User survey' (when auditors conduct a survey and/or perform user interviews to gather descriptive data of user experience on the platform);
3. 'Scraping audit' (when auditors collect data directly from a platform, typically by writing code to automatically click or scroll through a webpage to collect data of interest);
4. 'API audit' (when auditors access data through a programmatic interface provided by the platform that allows them to write computer programs to send and receive information to/from a platform);
5. 'Sock-puppet audit' (when auditors use computer programs to impersonate users on the platform and the data generated is recorded and analysed);
6. 'Crowd-sourced audit' (when real users are used to collect information from the platform during use – either by manually reporting their experience, or through automated means like a browser extension) (see **Ada Lovelace, 2021**).

Regulators need capacity, resources and skills to conduct these audits. In addition, civil-society and academic actors should be enabled to conduct these audits.

Algorithmic Debiasing. Furthermore, there is also the opportunity for debiasing algorithmic systems, especially in content moderation and search engines. Debiasing generally refers to "the application of select methods to address bias by achieving certain forms of statistical parity (**EDRi, 2021, p.23**)." As I have shown in Chapter 2, search engines ranking can indeed be biased. **Kay et al. (2015)**

show, for instance, undesired biases in the Google image search functionality when querying occupation-related images, with systematic underrepresentation of women and stereotype exaggeration. In cases like this, it is possible to 'debias' the algorithmic system. However, over-relying on algorithmic debiasing can represent a narrow approach that squeezes complex socio-technical problems into the domain of design and thus into the hands of technology companies. Effective solutions would actually require bold regulations that target the root of power imbalances inherent to the pervasive deployment of AI driven systems.

Competition and Economic Incentives

For many reasons, digital markets – in particular search engines and social media markets – tend to be highly concentrated with fundamental entry barriers for potential competitors (**Scott Morton et al., 2019**).

These are in part due to certain characteristics of digital technology (i.e. network effects), but in part also due to behaviors of market participants-consumers create entry barriers with their behavioral biases and companies by engaging in various activities. The resulting concentrated market structures do not serve consumers.

Level the Playing Field. A fundamental policy approach to mitigate many of the challenges and concerns coming from the algorithmic public opinion is arguably levelling the "playing field" of information intermediaries online, particularly resizing mainstream platforms and supporting emerging innovative companies. We might indeed expect the marketplace to self- correct (to some extent) and for companies that offer tools that are, for example, more privacy and autonomy-preserving to gain a competitive advantage. Such policy could be done by breaking big tech companies (for instance Whatsapp from Facebook, and Youtube from Google) and, at the same time, preventing them from acquiring potential competitors. So far, users had only an illusory choice: take it or leave a bunch of platforms. The problem is indeed systemic. Consider how in the past 20 y ears the GAFAM (Google, Apple, Facebook, Amazon, Microsoft) collectively bought 1,000 firms, and 97% of these transactions have not been vetted by anyone, very few challenged, and zero blocked in the US and around the world as well (**Cabral, 2021**). The result is that these companies became ever more powerful while we missed the impact of competition in so many different ways. It is also true, however, that to the extent that competition regulation would fragment intermediary markets and disperse market power, it might make harder, not easier, for online content harms to be addressed (**Bunting, 2018**).

Business Model. Unless the successfulness of the business model based on advertisements is completely undermined, we can't expect the market to easily correct itself. Many scholars argued that this business model inevitably built on engagement and popularity lead to divisive, emotional content, because that is what algorithms tend to favor. So, what could be the alternative to the attention economy business model? The most common and viable potential solution is changing to a subscription-based model. This would allow businesses

to own the relationship with consumers. Yet, this would reduce access to fundamental, ordinary services and information for those who can't, or won't, subscribe. This potential exclusion is a serious challenge to buildup alternative and sustainable social media business models. Perhaps a more convincing solution may be found focusing on the economic incentives, namely ads.

Tax, Ban or Limit Online Ads. The AD industry should help sellers to understand customers and deliver them marketing messages that are more relevant, consistent and effective. In general, it is argued that this model helps to fund the press and other channels of expression. Critics, instead, argue that this model incentivizes sensationalistic journalism, clickbait and, overall, negatively affects the quality of the press (**Kingaby and Kaltheuner, 2020**). More often, misinformation and conspiracy theories are the product of this business model, not an accident. For many reasons – among them adblockers (plug-ins to block ads), outof-sights ads (ads that are rarely seen by users) and click-farms (fake ads views) – it can be even radically questioned the effectiveness (and thus returns) of most online ads (**Neumann et al., 2020**). Not only there is little evidence that constant tracking leads to more relevant ads, but a recent study showed how targeted advertising accounted for only a percent average increase in revenue. Of course, Google has argued that publishers would lose half their revenue or more if they stop using personalized advertising. Actually, online ads seem so over-valued that might even represent the next financial bubble (**Hwang, 2020**). A radical solution that critics advocate is to ban micro-targeted ads and to opt instead for contextual ads, not only for their unintended and undesirable consequences but also because they appear to be limitedly effective. In addition or in alternati ve, a compelling policy approach is proposed by the Nobel prize for economics **Paul Romer** (**2021**) who argues to enact a progressive, sufficiently aggressive tax on revenue from digital advertising. This could make the subscription model more attractive or, more simply, to make it more attractive for a large firm to create independent new ventures, and less attractive for it to grow via acquisitions. To date, there is no policy agenda supporting these latter solutions. In EU, however, there are currently more than 20 proposals for behavioral ads reform being considered by legislators working on the Digital Services Act and Digital Markets Act.

Sustain Alternative Social Media. Mainstream platforms' business model could also be challenged by rival ideas and competitors. These competitors must stand a chance, not only thanks to active antitrust enforcement but also by seriously taxing large media (and other) corporations and channeling the resulting income into alternative noncommercial social media (**Fuchs & Marisol, 2015**). Indeed, so far the history of alternative media is a history of enormous challenges, mainly because hearing alternative voices is ultimately a matter of money and political resources to afford visibility. To some extent, this challenge is considered in EU. A paradigmatic example is the European project Next Generation Internet (NGI)[12]. This has funded, among many other projects, PeerTube – a free, libre and federated video platform – with 50.000 Euros.[13] Obviously, it might be contested that the current funding

12 https://www.ngi.eu/.

13 https://framablog.org/2021/11/30/peertube-v4-more-power-to-help-you-present-yourvideos/.

is insufficient to scale up and compete with mainstream platforms.

Strenghten Interoperability. Another promising policy solution is to strenghten interoperability[14]. Digital interoperability enables apps, digital services and devices to work together, even if they are made by different providers. Well-known examples of interoperability include e-mail, and telephone voice and messaging services — you can send an e-mail or text message or call anyone else, regardless of the service providers, apps or devices you use. In contrast, mainstream social media services like Facebook or Twitter as well as many other services, tend to only support interactions within their own platforms (e.g. a Facebook user cannot follow someone else's Twitter feed or a Telegram user is blocked from joining a WhatsApp group). This requires consumers to use multiple applications and devices that are incompatible with each other. Interoperability is one of the basic principles on which the internet was built. By adopting open technical standards, it would be possible to break down 'walled gardens' controlled by a single company. Interoperability of digital services can stimulate innovation by allowing new operators to enter markets such as social media and messaging services. This in turn creates an incentive for all operators to innovate and provide new features. Currently, mainstream platforms can rely on their network effects without having to compete on the merits of their products and services. Mandatory interoperability for the largest digital platforms can also foster the creation of whole new digital markets where startups build digital services on top of incumbent platforms. That way, consumers can get access to better AI-driven content moderation algorithms that run on top of a user's Twitter feed or Facebook timeline. In the same vein, new apps could replace YouTube's recommender algorithm that is known to promote extremist video content even to people who weren't looking for it.

Data Protection and Privacy

Algorithmic systems have the potential to transform seemingly non-sensitive data into sensitive data about individuals. At times, such transformations can create the possibility for discrimination against individuals and groups, or simply violate expectations and create data flows and knowledge that individuals may find inappropriate. A range of issues surround the transformation of data from sensors or online behaviour into sensitive data that concerns an individual's health, wellbeing or mental state.

Data protection is, of course, essential for privacy. It is fundamental not only to avoid possible manipulations of the algorithmic public opinion but also for individual and collective human rights. Of course in this section I do not intend to raise all the important issues concerning privacy and data security that relate to the algorithmic public opinion. There are several other efforts underway that aim to be more comprehensive than this one, for example on the data ecosystem and governance (e.g. **World Bank, 2021**), the debate on data sovereignty (e.g. **De La Chapelle and Porciuncula, 2021**), data ownership and the future of data (e.g. **Decode, 2019; Mills, 2020; Ada Lovelace Institute,**

14 For more information on interoperability see https://interoperability.news/.

2021). I hope, however, to shed lights on a number of open technical and policy questions and debates such as: how to overcome the limitations of platforms' informed consent? When dark patterns are manipulative should they be outlawed? And what new empowering rights need data consumers in this context? How to build a new paradigm for data ownership in this context?

Protect Group Privacy. Due to a set of externalities and information leakages involved in data markets, privacy could be actually recognized as an "aggregate public good" prone to market failure. Recognizing this should convince us that government intervention is both beneficial and necessary for its protection. This is even more relevant considering the flaws of privacy self-management and emerging algorithmic techniques that might threaten the privacy of groups. In fact, algorithmic systems which measure, count, and profile groups of individuals create knowledge that is not (only) private to an individual, but which reveals something about a group of individuals. This is framed as the protection of group privacy. Since long privacy researchers proposed a contextual and relational understanding of privacy, mainly refferred to as relational privacy. The main problem is that algorithms can create ad hoc and temporary groups to which none of the existing interpretations of privacy can be applied. The implication of the technical issues related to group privacy is that our legal, philosophical, and analytic attention to the individual may need to be adjusted, and possibly extended (**Taylor et al. 2016**). The fact that the individual is no longer central, but incidental to these types of processes, challenges the very foundations of most Western legal, ethical, and social practices and theories related to privacy.

Users' Data as a Public Resource? As regulatory framework, **Napoli (2019)** proposes to treat users' data as a public resource. The central premise is that, whatever the exact nature of one's individual property rights in one's user data may be, when these data are aggregated across millions of users, their fundamental character changes in such a way that they are best conceptualized as a public resource. Certainly, it is in this massive aggregation that the economic value of user data emerges. Therefore, if policymakers would treat aggregate user data as a public resource akin to broadcast spectrum, that framework would provide what may be the most constitutionally robust rationale for imposing public interest obligations upon those social media platforms that rely upon the aggregation and monetization of user data. This approach is legally grounded as well as promising for it can allow researchers and civil society to access and analyze social big data. In EU, the Art. 31 of the proposed Digital Services Act would give researchers with academic affiliations access to platform data for public interest research (see **Leerssen, 2021**).

Beyond Individual Consent? The individualistic approach of privacy self-management usually relies on informed consent, but this seems not be an optimal solution because it leads to uncertainty and context dependence. People in fact cannot be counted on to navigate the complex trade-offs involving privacy self-management (**Acquisti et al.,**

2015). Most people in fact neither read nor understand online privacy policies. These have also two inherently contradictory goals: to be understandable to consumers – which requires simplicity and brevity – and say something meaningful about how data is processed – which is complicated and requires a lot of details. As well known, it would take an enormous amount of time to read all the conditions of the websites we visit – 201 hours on average per year. Moreover, dark patterns are often employed during terms of conditions and privacy updates to nudge consumers toward options that benefit company profitability but may not reflect consumers' actual preferences or expectations (**Moen et al., 2018**). Even worse, the least educated seem most likely to be manipulated successfully (**Luguri and Strahilevitz, 2021**).

As **Strahilevitz et al. 2021** argue, a compelling solution to this model is that the content of contractual default provisions would depend on the articulated preferences of ordinary consumers as measured by scientifically rigorous survey instruments. In privacy and security settings there are in fact many instances in which it is appropriate for the law to use "consumertarian" default rules – i.e., the legal defaults preferred or expected by a majority of consumers. This would arguably alling users' preferences to platforms' terms and conditions. Of course, such an approach is not devoid of implementation challenges. More generally, nonetheless, the development of informed consent with respect to big data use is highly debated and other proposals are indeed discussed (see **Andreotta et al., 2021**).

A Ban to Dark Patterns? Many of the inherent problems with dark patterns have implications for information privacy. Dark patterns are indeed often used to direct users toward outcomes that involve greater data collection and processing. Additionally, the proliferation of data-driven computational methods allows firms to identify vulnerabilities of users and to target specific users with these vulnerabilities. While dark patterns come in a variety of different forms, their central unifying feature is that they are manipulative, rather than persuasive. More specifically, the design choices inherent in dark patterns push users towards specific actions without valid appeals to emotion or reason. The line between manipulation and persuasion, however, is sometimes difficult to draw, not only ethically but also legally. **Strahilevitz et al. 2021** propose a framework that could allow legislators, regulators, and courts to define the category of manipulations warranting legal action in a way that is workable and defensible on both economic and moral grounds.

Editorial Obligations and Intermediary Liability

Considering the special position of media and the preference for self-regulation in this area, for some time it has been considered unlikely that the normative principles in the media context would directly translate into legal obligations. Recent debate on the public responsibility of social media platforms pivots on the question of whether or not platforms can be held accountable for the content shared through them, legally and morally. Platforms indeed still enjoy the same

status as Internet Service Providers (ISPs) under US (Communications Decency Act, 1996) and EU law (E-COMMERCE Directive 2000/31/EC). The growth of information intermediaries, and their significant control over distribution of certain kinds of content, has in fact reinvigorated public debate about the appropriate balance between competing rights to free expression and protection from harmful or illegal content. Today there is an increasing consensus - especially in the European Union - that due to market concentration new legal approaches and new governance models have become necessary. New definitions seem to be needed to address the role of the information intermediaries and distinguish them from ISPs. Mainstream platforms are difficult to frame, and their consequences are equally difficult to govern and regulate.

The actual role and capacities of social media's platforms to prevent certain undesirable outcomes or to contribute to their realization is still debated. Of course, platforms fundamentally shape user activity, yet they do not determine this activity. Many of the problems with media pluralism and diversity are, to a large extent, user-driven. For similar reasons, at least part of the solution to potential public policy challenges lies with the users. The current focus in law on allocating responsibility to one central party – editor, data controller, or the supplier of a service – is primarily due to the fact that this central actor is the source of potential risk or harm, or the controller of a resource that can give rise to legal responsibilities. Yet, multiple actors are effectively responsible.

From a legal point of view, this discussion is grounded in the host-editor, namely either social media qualify as hosts, with the consequence that they fall under the European e-Commerce regime, or they are categorized as editors, having full legal responsibility for what is shared through their platforms. As many scholars have argued, the legally enshrined conceptual framing of a "platform" that merely hosts content, but should not be held legally liable for it, became a strategic and powerful enabler for the rise of today's digital giants. Ample scholarship has since shown that the framing of technology companies as mere "hosts" or "intermediaries" or "platforms" elides the ways in which these companies set norms around content or speech , algorithmically recommend content, and assume a host of functions that combine features of publishers, media companies, telecommunications providers, and other firms (**Napoli and Caplan, 2017**).

Yet, there are three specific problems with imposing editorial obligations on intermediaries:

1. The lack of accountability. Editorial obligations tend to increase intermediaries' power over content markets, without necessarily increasing the transparency or accountability of its use. Large volumes of material may be removed or blocked, with little visibility of the true social benefits and costs.
2. The impact on competition. Making intermediaries responsible for the content they host may have a disproportionate impact on new entrants for whom the cost of preventing content infringements

would represent a substantial burden. Most intermediaries process a vast volume of content, much greater than any traditional publisher, meaning that the costs of proactive monitoring of content are also much greater. Yet, regulation can refer to specific requirements that platforms should have to avoid to burden emerging platforms.

3. Imposing editorial obligations on intermediaries may have limited effect. Harmful content can easily flow from more regulated to less regulated environments, where it may be less visible and less susceptible to responsible intermediary activity. For example, aften the removal of Trump on Facebook and Twitter millions of users moved to the right-wing platform Gab.

Editorial obligations skew intermediary incentives towards those interests which are protected by takedown; the stronger the incentives on intermediaries to remove content rapidly, the greater the likelihood of legal content being inadvertently blocked. These cases raise questions of fundamental rights, and how rights that may be in tension are to be balanced. The goal of blocking illegal content must be reconciled with the risk of inadvertent denial of access to legal content. But editorial obligations and liabilities create no incentive for intermediaries to consider a 'fair' or 'just' balance, only to secure the commercially optimal outcome, which will skew towards content takedown in proportion to the size of the sanction for distributing illegal content. The existing content regulatory toolkit is wholly unsuited to the task. New ways of regulating intermediaries – which reconcile their market governance with protection and balancing of rights – are needed.

Specific Issues

To complement systemic issues, in this policy review I also focus on a number of more specific issues I have identified and that are paramount, namely recommender systems, content moderation, social bots, political microtargeting and, finally, disinformation. All these are certainly intertwined with the above systemic issues. To begin, the analysis of recommender systems is stricly related, in particular, to the governance of algorithms, and it is fundamental to develop personalized filtering that are fair and accountable. Similarly, political microtargeting has to do with algorithmically personalized political advertisement. Then, content moderation is equally relevant in the content management of social media platforms, while social bots are a concerning weapons of online manipulation, often disrupting the marketplace of ideas. Finally, disinformation is a more systemic issue that includes all of the previous issues and, in fact, has received more attention from public opinion and policy-makers.

Recommender Systems

Recommender systems (RSs) are probably the most relevant algorithmic system in the context of the algorithmic public opinion. In particular, RSs generate serious concerns on hate speech,

disinformation and conspiracy theories, but also on monopolisation and platform power. Legal responses to these problems are not straightforward given the various stakeholders and the amount of information that platforms typically deal with. Any form of restriction on recommendations is difficult to automate, culturally contextual, and potentially sensitive. Also, regulation focusing on the transmission or hosting of content itself brings freedom of expression concerns. Yet, as Cobbe and Singh **(2019)** argue, the same risks do not necessarily arise from regulating the further dissemination of content by platforms. While the fundamental right of freedom of expression should be respected as far as possible, individuals do not have a fundamental right to have their speech disseminated or amplified by platforms in this way (as the adagio goes "freedom of speech is not freedom of reach"). By focusing on recommending, rather than on the transmission or hosting of content itself, regulation can largely sidestep these freedom of expression problems and focus on the use of technical systems by private corporations to pursue their own business goals.

For recommender systems to be 'responsible', Cobbe and Singh (2019) outlined a number of principles for the service providers who use these systems, particularly (open) RSs systems[15]. In short, they must be lawful and service providers should be prohibited from doing it where they violate these principles. Indeed, if service providers can't use RSs responsibly then they shouldn't be permitted to do it at all. Similarly, service providers should have conditional liability. No liability protection would therefore be available when undertaking a prohibited practice. Service providers should indeed have a responsibility to not recommend certain 'potentially problematic content'. This should establish a responsibility to not promote certain kinds of content (for example, white supremacism, health disinformation, anti-Semitic conspiracy theories, pro-suicide or self-harm content, content promoting eating disorders, and so on). Any potential liability would therefore actually result from undertaking a prohibited practice rather than from recommending certain kinds of potentially problematic but lawful content.

Individual Control. First of all, RSs should be opt-in, meaning users should be able to exercise a minimum level of control over recommending, and opting-out again should be easy. Offering control to users is a good idea if, and only if, those who do not exercise it do not end up being treated less favourably than those who do. To this end, it would be desirable for RSs to be available to users only on an opt-in basis. Users who choose to receive recommendations should, at a minimum, be able to: exclude certain content from recommendations, exclude certain sources of content from recommendations, exclude certain of their behaviours or interests from the process of determining what should be recommended to them, and to easily and freely opt back out of recommendations entirely. Furthermore, users could be enabled to choose between different RSs, including some from third parties.

User-facing disclosures. Similarly, user-facing disclosures aim to

15 To be more specific, in fact, the following principles refer to 'open recommenders' which are those that provides recommendations from a pool of content which is primarily user-generated or submitted, brought in automatically from various sources, or otherwise aggregated in some way without being specifically selected by the platform. Examples include Google, YouTube, Facebook, Reddit, Instagram and Amazon. These differentiate from 'curated recommending' in which the system recommends from a pool of content which is curated, approved, or otherwise chosen by the platform rather than provided directly by users or advertisers or automatically brought in from elsewhere and do not typically include user-generated content without some kind of editorial process (Netflix is a popular example of a curated system). Finally, they are distinguished from 'closed recommending' where the content to be recommended is generated by the platform itself or the organisation which operates that site. For example, where a news organisation provides a personalised feed of stories and articles to its users, all of which are produced or commissioned by the organisation itself.

channel information towards individual users in order to empower them in relationship to RSs (e.g. Facebook's 'Why Am I Seeing This' feature). The aim of such transparency is to inform users about their available options so as to help them form their own preferences, appealing to values such as individual autonomy, agency and trust.

Public Disclosures. Another significant transparency requirement is that service providers should be required to keep records and make information about recommendations available to help inform users and facilitate oversight (i.e. public disclosures). The constantly changing nature of social media and other online services makes it difficult to identify and track problems over time. Service providers should, at a minimum, be required to keep logs of recommended content (both for personalisation and for behavioural targeting) so that they can be reviewed by users and by oversight bodies (for a discussion see **Leerssen, 2020**). Of course, provided that these should have privacy-by-design and a trustless design that pre-empts abuse by malicious actors, public records could be instrumental for purposes of real-time, high-level monitoring by media watchdogs such as journalists, activists, and NGOs. These may not suffice to conclusively demonstrate bias or discrimination in RSs, but at a minimum they could offer a starting point for such investigations and serve as a first-warning system for more targeted efforts.

New Oversight Authority. Government oversight could appoint a public entity to monitor RSs for compliance with publicly regulated standards. This endeavior faces many significant challenges, both practical and principled. To begin, government authorities are capacity-constrained, particularly with regard to the technical expertise required to perform complex algorithmic auditing. This is especially true for horizontal agencies such as competition and data protection authorities, for whom RSs risk being overshadowed and overlooked. Sectoral proposals, instead, would in many cases require the creation of entirely new oversight bodies. Government auditing powers also raise issues of 'second-order accountability': is the governance system itself sufficiently open to outside scrutiny? If government determinations rely on privileged access to confidential data, which is not accessible to broader publics, it may be difficult for citizens to fact-check and second-guess government policy in this space. Without broader forms of second-order transparency and accountability, the legitimacy of a technocratic, command-and-control approach in such a politically sensitive, value-laden context can therefore be called into question.

Access to Data for Research. Partnerships with academia and civil society would better enable these stakeholders to research and critique RSs. Yet, these are often met with skepticism for several reasons. Above all, creating meaningful transparency arguably runs counter to platforms' incentives: they have a commercial interest in monetizing traffic data and insights, and thus in keeping this information exclusive, and a political interest in avoiding negative publicity. Besides independent surveying, one of the most important sources of data regarding recommender systems has been their public APIs, through

which outside researchers can download platform data in bulk. But these have come under significant pressure over the past years. In the EU, it is finally discussed the regulation of researcher access to platform data. Article 31 of the proposed Digital Services Act (DSA) on "Data Access and Scrutiny" is the first legislative framework for researcher access to platform data. It has been welcomed with a mix of hopes and criticisms (**Leerssen, 2021**).

Other Restrictions. There should be also specific restrictions on service providers' ability to use recommending to influence markets through RSs. Service providers should be explicitly prohibited from unduly recommending their own products and services ahead of those offered by others. These prohibitions would not only complement and refine the existing data protection principle of purpose limitation, but would go some way towards addressing competition issues arising from the dominance of platforms and their use of personal data, particularly where leveraging their dominant position in one market to gain a competitive advantage in another. Beyond these general principles for 'responsible RSs', other applicable legal frameworks must also be considered. Data protection, in particular, is fundamental in this context, given the extensive behavioural tracking and processing of personal data which underpins RSs. Similarly, service providers cannot ignore their responsibilities under equality and non-discrimination law or, indeed, any other applicable regime.

Content Moderation

Content moderation not only involves algorithms but also users in conducting review processes, which requires them to set content standards that are able to be easily encoded, interpreted and applied consistently across widely varying national jurisdictions. This would suggest that platforms engage in moderation similarly to the editorial and governance processes undertaken by legacy media companies in content regulation. Yet, the user led nature of content generation, the scale of creation, the inclusion of community reporting and the decentralization of decision-making across time zones and cultural contexts makes these processes more diverse, complex and demanding to negotiate than traditional, professionally oriented editorial decisions.

The massive user generated flows of digital platforms today have been built on the premise that content can be post moderated – although AI-oriented automation will eventually enable high degrees of algorithmic pre-moderation. However, there is a large disjuncture, according to Gillespie **(2018)**, between the 'data scale' of faceless, consistency oriented, automated regulation and the 'human scale' of localized, culturally bound interactions, which suggests that machine controls will never be a complete solution to the challenge of classifying endless content flows. Addressing the problem of national differences in content regulation and cultural expectations of publishers is also key to the future of platform governance. At the same time, there is a myth that these systems are completely automated: there is always a decision-making process behind categorising content, behaviours and people as

deviant, but so far these have been hidden and unaccountable.

Furthermore, public opinion have mostly focused on the latest controversies and the biggest players. Major platforms are in fact enormous and their policies affect billions of users. Their size makes them desirable venues for malicous actors. Their policies and techniques set a standard for how content moderation works on other platforms. Yet, the largest, US-based platforms do not provide a reliable guide for the entire social media ecology. There are many kinds of social media platforms that configure content moderation differently. Moderation also happens on sites and services different to mainstream social media: on comment threads and discussion forums, in multi-player game worlds, in app stores, on dating sites, and on the many other services. These also differ in ways that affect how content moderation works: by size, reach, and language, but also by technical design, genre, corporate ethos, business model, and stated purpose.

The content moderation debate should expand beyond treating platforms as primarily venues for public speech, or as silos that exist in isolation from one another (**Gillespie et al., 2020**). Instead, we might think of them as a web of private infrastructures that we traverse in our digitally mediated lives. They are indeed also marketplaces, payment systems, advertisers, gaming sites, and media distributors. Rather than thinking about content moderation in terms of its effects on speech alone, we should instead consider the consequences that moderation can have on communities, by influencing the access to platforms that are increasingly central to our ability to work, live, and socialize.

There are also a number of critical areas to improve moderation and minimize its risks. Innovations in automated content moderation have focused overwhelmingly on identification techniques such as detect pornography, harassment, or hate speech. Not only are there problems with these ambitions but automated content analysis tools have shadowed other possible uses of algorithms to support content moderation. Research should also prioritise tools that might support human moderators, community managers, and individual users to eventually make more informed decisions and data-scientific techniques might also help users and community managers better grasp how differently other communities experience similar content or behaviour.

Another underdiscussed but significant issue is the recording of the content that has been removed. In particular, social media documentation of human rights violations is critical for justice and accountability efforts, and in some cases it serves as collective memory. Videos and text posted online are living histories for some diaspora communities, and sometimes this documentation might offer the only evidence that a crime has been committed. Yet in too many cases, social media content moderation policies around extremism lead to the deletion of vital documentation. Restoring wrongfully deleted content is nearly impossible if the person who posted the content is not alive, is arrested, or does not have access to email, all common issues in conflict zones.

Nonetheless, as much as platform content moderation could improve, it may also be a perennially impossible task to do in such a way that no one encounters harm, friction, or restriction. Users of social media may have unreasonably high hopes for what their experience should be, largely because of the endless promises made by social media platforms that it would be so. We need to educate and adjust the expectations of users, to both understand what a difficult and vital process this is, to demand it be transparent and accountable, to recognise how they are implicated in it, and to prod their sense of agency and ownership of these sometimes unavoidable dilemmas.

Social Bots

Gorwa and Guilbeault **(2019)** argue that multiple forms of ambiguity are responsible for much of the complexity underlying contemporary bot-related policy, and that before successful policy interventions can be formulated, a more comprehensive understanding of bots—especially how they are defined and measured—will be needed.

Any initiatives suggested by policymakers and informed by research will have to deal with several pressing challenges: the conceptual ambiguity (highlighted in the Chapter 1), poor measurement and data access, lack of clarity about who exactly is responsible, and the overarching challenge of business incentives that are not predisposed toward resolving the aforementioned issues.

Measurement and Data Access. Bot detection is very difficult. Researchers are in fact unable to fully represent the scale of the current issue by relying solely on data provided through public APIs. Notoriously, they cannot study bots on Facebook and virtually all studies of bot activity have taken place on Twitter which already questions whether their APIs provide a fair account of content on the platform. Even the social media companies themselves find bot detection a challenge, partially because of the massive scale on which they function. Tracking the thousands of bot accounts created every day, when maintaining a totally open API, is virtually impossible. Taking this a step further by trying to link malicious activity to a specific actor (e.g., groups linked to a foreign government) is even more difficult, as Internet Protocol (IP) addresses and other indicators can be easily manipulated.

Even the most advanced current bot detection methods hinge on the successful identification of bot accounts by human beings. The problem is indeed that humans are not particularly good at identifying bot accounts. Researchers can never be 100 percent certain that an account is truly a bot, posing a challenge for machine learning models that use human-labeled training data. The precision and recall of academic bot detection methods, while constantly improving is still seriously limited. Of course, less is known about the detection methods deployed by the private sector and contracted by government agencies, but one can assume that they suffer from the same issues. Just like researchers, governments have data access challenges. This pose substantial challenges to identify the scale of bot activity, epsecially during

elections. The policy implications of these measurement challenges become very apparent in the context of the recent debate over the role of Russian propaganda during the 2016 U.S. Presidential election. To understand the scope and scale of the problem, policymakers will need more reliable indicators and better measurements than are currently available.

Responsibility. A key, and unresolved challenge for policy is the question of responsibility, and the interrelated questions of jurisdiction and authority. To what extent should social media companies be held responsible for the dealings of social bots? And who will hold these companies to account? A whole spectrum of regulatory options under this umbrella exist, with some being particularly troubling. For example, some have argued that the answer to the "bot problem" is as simple as implementing and enforcing strict "real-name" policies on Twitter—and making these policies stricter for Facebook. The recent emergence of bots into the public discourse has reopened age old debates about anonymity and privacy online, now with the added challenge of balancing the anonymity that can be abused by sockpuppets and automated fake accounts, and the anonymity that empowers activists and promotes free speech around the world. In a sense, technology companies have already admitted at least some degree of responsibility for the current political impact of the misinformation ecosystem, within which bots play an important role. The matter is by no means settled, and will play an important part in the deeper public and scholarly conversation around key issues of platform responsibility, governance, and accountability.

Contrasting Incentives. Underlying these challenges is a more fundamental question about the business models and incentives of social media companies. Business incentives are indeed critical in shaping content policy—and therefore policies concerning automation—for social media companies, slightly different incentives have yielded differing policies on automation and content. Notably, Twitter's core concern has been to increase their traffic and to maintain as open a platform as possible. Thus, it allows to use tools that automate their activity, which can be useful: accounts run by media organizations, for example, can automatically tweet every time a new article is published but also fulfill many creative, productive functions. Facebook, instead, has been battling invasive spam for years and has much tighter controls over its API. As such, it appears that Facebook has comparatively much lower numbers of automated users (both proportionally and absolutely) than Twitter, but is concerned primarily with manually controlled sockpuppet accounts, which can be set up by anyone and are difficult or impossible to detect if they do not coordinate at scale or draw too much attention. Thus, platform interests often clash with the preferences of the academic research community and of the public. Academics strive to open the black box and better understand the role that bots play in public debate and information diffusion, while pushing for greater transparency and more access to the relevant data, with little concern for the business dealings of a social networking platform.

There are no easy solutions to these challenges, given the complex tradeoffs and differing stakeholder incentives at play. As a highly political, topical, and important technology policy issue, the question of political automation raises a number fundamental questions about platform responsibility and governance that have yet to be fully explored by scholars. Conceptual ambiguity can be reduced by diligent scholarship, and researchers can work to improve detection models, but responsibility may not be taken and business incentives will not shift on their own. Despite mounting concern about digital influence operations through social bots over social media, especially from foreign sources, there have yet to be any governmental policy interventions developed to more closely manage their political uses. An important measure has been proposed by state legislators in California in April 2018. This was requiring that all detected bot accounts were publicly labeled by social media companies.

Political Micro-targeting

Political micro-targeting (or 'behavioral advertising', 'political ads' or even 'dark ads') is defined as 'creating finely honed messages targeted at narrow categories of voters' based on data analysis 'garnered from individuals' demographic characteristics and consumer and lifestyle habits'. It can be summarised as consisting of three steps: 1) collecting personal data, 2) using those data to identify groups of people that are likely susceptible to a certain message, and 3) sending tailored online messages (**Zuiderveen Borgesius et al., 2018**). As such, unlike traditional political advertising, micro-targeting not only affects the democratic process, but it also affects people's privacy and data protection rights. Indeed, micro-targeting affects myriad other rights and duties, including a political party's and online platform's right to impart information, a voter's right to receive information, and the government's duty to ensure free and fair elections.

The objectives of political micro-targeting can be manifold: to persuade, inform, or mobilise, or rather to dissuade, confuse or demobilise voters. It can conceivably be used in interesting ways to be part of government communication and thereby unfold not only as a feature of campaigning but also of governing. People can be micro-targeted on the basis of all kinds of information (such as their personality traits, their location, or the issues they care about). Hence, any data can be valuable: from consumer data to browsing behaviour. Such data can provide enough information to make inferences about the susceptibilities of the target audiences. A micro-targeted audience indeed receives a message tailored to one or several specific characteristic(s). This characteristic is perceived by the political advertiser as instrumental in making the audience member susceptible to that tailored message. For example, when micro-targeting a party could ignore the unlikely voters and tailor their messages to possible voters' issue salience (or other characteristics). During the Brexit referendum, the cross-party "Vote Leave" campaign commissioned well 1,433 customized adverts promoting a more or less explicit pro-Brexit message (**Reviglio & Agosti, 2020**). A regular targeted message, instead,

does not really consider matters of audience heterogeneity.

Micro-targeting originates from the United States, where there are relatively loose data-protection. It seems likely that, when compared to the US, Europe's privacy rules hinder micro-targeting. Anyway, national parliaments seem best placed to regulate political micro-targeting, especially for what concerns electoral laws.

There are several concerns related to political microtargeted ads. Despite changes that were introduced by most platforms, political advertising continues to lack the transparency necessary to ensure fair and democratic elections. Many changes introduced by the platforms haven't been rolled out globally, meaning countries with volatile political contexts and fragile democracies risk being most vulnerable to election interference. According to a 2019 study by Privacy International, Facebook only required political advertisers to be authorised, or for political ads to carry disclosures, in around 17% of countries around the world. Google provides 'heightened transparency' for political ads in 30 countries – around 15%. Furthermore, political microtargeting techniques can also amplify the effects of deepfakes, but for a much smaller subgroup than expected (**Dobber et al., 2019**). Deepfakes can indeed poison the public debate by confusing people on what is real and what is not. A number of counteractions might certainly mitigate the challenges that political micro-targeting poses.

(see → Box 4 ***The Rise of Deepfakes***)

Conceptual clarity. Definitions of political ads vary widely. Thus, a consensus is needed, at least in specific contexts.

More Research. We still need to understand how benefits of microtargeting outweigh the risks. Therefore, research is needed, not only to bring the above conceptual clarity but also to assess the effectiveness of these techniques as well as help to develop ways to scrutinize these delicate processes of public opinion formation. More generally, it is fundamental both to reach more transparency and to access relevant data.

More transparency. Necessary but not sufficient, there are a number of potential transparency requirements: from expenditures of political parties and from intermediaries' ads, we need to set a regulatory framework that makes 'ads central repositories' accountable. It is possible to legislate so that all paid-for political adverts can be viewed by the public. Yet, when it comes to such political ad libraries, there is a lack of standardisation. In any case, these should be able to inform users to see if:

1. A political advertiser microtargeting;
2. A true, verified name of the advertiser in the disclaimer about who paid for the ad.
3. Eventually require political parties to disclose their campaign finances broken down by media outlet.
4. How a platform has amplified the ad.

5. Give an existing body the power to regulate political advertising content or create a new one to do so.
6. require all factual claims used in political adverts to be independently substantiated.

Limit political ads usage. Policy-makers could probably ban online political micro-targeting, at least for a period leading up to elections and during the run-up like referendum. Many countries have already sector-specific rules for political advertising, which differ from country to country. For decades, paid political advertising on television has been completely banned during elections in many European democracies. These political advertising bans aim to prevent the distortion of the democratic process by financially powerful interests, and to ensure a level playing field during elections. Member States' rules for political television advertising still differ widely from what is required for online political advertising. Policy-makers could nsure that rules for online political advertising match those for offline political advertising on traditional media in the context of local, national or sovra-nationals (e.g. EU) elections.

Limit psychometric profiles. Psychometric profiles could be labeled a 'special category' of data. Citizens seem to react differently to affect-based political ads based on their psychometric profile: introverted people generated higher voting intentions when they were targeted with a negative fear-based political ad, whereas extraverted citizens had higher voting intentions after receiving a positive enthusiasm-based political ad. This means that the processing of data revealing a person's psychometric profile is only allowed under specific conditions, such as receiving a person's 'informed consent'. Either, an outright ban on psychometric profile can be considered.

Disinformation

Production of misleading or false content is not limited to the current historical context, but has always been inherent in human communication. In the algorithmic public opinion, however, scale, consequences and responses are arguably more complex and various. Many recommendations have been provided for this particular issue which is at the intersection of the systemic and specific issues hihglighted in this policy review.

Any interventions in the online disinformation space have to recognise that domestic media and domestic politicians are often part of some disinformation problems – and, importantly, the public recognises this and frequently expresses the same level of concern over what they see as political propaganda and poor journalism as they do over false and fabricated content (**Newman et al. 2018**).

Importantly, responses should be fully compliant with the fundamental principles of freedom of expression, free press and pluralism, and at the same time future-proof and efficient in averting public harm. In order to ensure this, the independent High level

Group on fake news and online disinformation (**de Cock Buning, 2018**) cautioned against simplistic solutions. More generally, it is recommended for a multi-dimensional approach that caters for the need to continually examine the phenomenon and evaluate the effectiveness of the concrete measures adopted by different actors. At the same time, it is advocated a self-regulatory approach based on a clearly defined multi-stakeholder engagement process, framed within a binding roadmap for implementation, and focused on a set of short and medium-term actions. There is considerable scope for expanding and improving the collaborative approach to combating disinformation by involving all relevant stakeholders (public authorities, platform companies, private news media, public-service media, and civil society groups, including factcheckers, media literacy groups, and researchers).

Potential and enacted policies are various and employed at different levels, actors and part of the process (**de Cock Buning, 2018**); More generally, these can be summarized as:

1. Enhance transparency of the digital information ecosystem;
2. Promote and sharpen the use of media and information literacy approaches to counter disinformation and help users navigate our digital information environment;
3. Develop tools for empowering users and journalists and foster a positive engagement with fast-evolving information technologies;
4. Safeguard the diversity and sustainability of the news media ecosystem, and, finally
5. Calibrate the effectiveness of the responses through continuous research on the impact of disinformation and an engagement process that includes predefined and time-framed steps combined with monitoring and reporting requirements.

One of the main concerns concerning disinformation is foreign influence and propaganda, especially from Russia which has allegedly used cyber-attacks, disinformation, and financial influence campaigns to meddle with the internal affairs of most European countries in order to amplify existing political and social discord and erode trust in mainstream media and democratic institutions. In particular, elections have become a target for hybrid-warfare. Although the electoral voting systems have not been compromised yet, attacks against auxiliary services have been attempted in some countries. To deal with these troubling forms of disinformation, there is an urgent need for an official coordinating body to enable collaboration between private companies and democratic governments, allowing them to work together to identify and thwart foreign information operations.

Then, while slow, expensive, and limited in scope, significant investment in media literacy for citizens of all ages is also likely to be a key part of increasing societal resilience to various kinds of disinformation. To make a meaningful difference, media literacy has

to be a central part of education (as it is already is in some countries) and significant resources will have to be invested in media literacy for adults, as a growing body of research suggests that older people may be both more exposed to disinformation and more likely to share it. For media and information literacy to be effective, it must be pursued across teacher training curricula, school curricula, and beyond, and it will require significant investment and ongoing evaluation and evolution.

Finally, to develop credible and effective policy responses to disinformation, there is an urgent need for more independent, evidence-based research. While there is no doubt that there are many and serious problems of disinformation, we still know little about the scale and scope in different countries, the actual effects of disinformation, and the effectiveness of various possible policy interventions.

Conclusions

This policy review has provided an overview of the challenges of the algorithmic public opinion and the debate on proposed policy approaches and promising proposals. The algorithmic public opinion is a complex infrastructure that comprises various algorithms and these algorithms, in turn, are not discrete tools that inscribe values and follow specific directives that can be regulated as such, but socio-technical assemblages that include a variety of actors, stakeholders, processes and technologies. Ultimately, algorithmic systems represent only a point of departure to understand the complexity of how public opinion is disrupted by this infrastructure. To regulate the algorithmic public opinion means to regulate the environment in which algorithms perform; this means to understand why they are employed, what are the ideologies that justify their usage, how algorithms operate in specific contexts and how they change and are adapted over time, what data are used for algorithms and how, how users interact with algorithm, how these are visually represented in interface design, how they are generally understood, and what are their (unintended) consequences. Still, the definition of these algorithmic systems is fundamental for an effective regulation but these are still famously blurred in academia as well as in policy-making and common usage.

In this policy review, I have analyzed the most significant algorithmic systems molding the algorithmic public opinion. Arguably, the most influencing ones are platform recommender systems, the ones that act as fundamental global gatekeepers allowing content personalization. The gatekeeping role of personalization algorithms sets up an unprecedented immense 'opinion power' that represents the most concerning regulatory challenge. Yet, other relevant algorithms have been discussed; the ones employed in content moderation (to detect and remove content), social bots (to manipulate content diffusion and discussions), and, among others, those who help to create, edit and retrieve content.

The regulation and governance of an algorithmic public opinion is certainly unprecedented. This is especially true considering not only the intrinsic features that make difficult algorithms to research and, thus, to understand, but also because the data that researchers can currently access is substantially limited and more often unreliable. Another important - perhaps obvious - conclusion of this policy review, in fact, is that it is paramount to allow researchers to access wider dataset in a legal and transparent manner. This is a *condition sine qua non* without which it becomes very difficult to collect convincing evidence for effective policies as well as to maintain a critical scrutiny over these powerful algorithms by civil society. The current institutional barriers to the social data produced in online platforms is indeed a form of *agnotology* – the science of ignorance – and the ignorance it produces, the often ambivalent and constrasting conclusions that scholars draw from limited datasets, seriously hinders the ability to govern and regulate this fundamental infrastructure of information societies, while it has created a perverse dynamic in which researchers could get more funding and platforms could escape regulation.

The risks that an unregulated algorithmic public opinion pose are undoubtedly grave, even if scientific evidences are often unconvincing. These include not only threats to democratic processes and values, including electoral integrity, but also, more concretely, misinformation, political polarization, radicalization, and even "addiction". These can ultimately lead to collective distraction, confusion and disagreement that can conduce to dangerous collective inaction in front of epochal challenges, Covid-19 and climate change. For the most critical, this could represent an epistemic (and democratic) implosion; in fact, the pollution of public opinion can lead to public preferences different to if these were accurately informed, which can have negative policy implications. The same is true with public opinion more generally, where policy outputs feed back on public inputs into the policy-making process. Therefore, policies for the algorithmic public opinion can even increase the quality of public policies more generally. This makes their effectiveness ever more significant.

The question of regulating the algorithmic public opinion is part of a complex regulatory system involving many different legislative acts and normative approaches. Policy-making, however, still lacks a shared vocabulary or frameworks for approaching the incoming challenges and new models for digital governance will likely need to be developed. The algorithmic public opinion is particularly challenging to policy-makers as it intersects with a transformation of media itself, due to media convergence, media globalization and the rise of global digital platforms. Legislation in the algorithmic public opinion context always risks being too broad to be useful, or too specific to be future-proof.

Despite several regulatory attempts in various legal systems, current regulation and policies are generally still unable to provide adequate responses to the challenges that the algorithmic public opinion poses. Rules and policies are fragmented and mostly provide horizontal approaches to many of these challenges, that is, diverse regulatory areas

concur to regulate algorithms ultimately reducing their effectiveness. More generally, the systemic policy challenges analyzed in this policy review stress that the Internet is a global common infrastructure rooted in all societies' economies, politics, and cultures, and that the tensions it produces are not promptly solvable by the current national and international legal system. As such, there is the need for a trans-national governance, especially because the worldwide competing policy approaches may lead to the fragmentation and, eventually, cyber-balkanization of the Internet. This realization should lead us to radically re-think the governance of the Internet itself. At the same time, we should also avoid naive expectations of the Internet as an ideal public sphere: there are always new challenges, and human and machines mistakes will inevitably raise further concerns. It is a never-ending challenge the governance of the algorithmic public opinion.

Bibliography

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347(6221), 509-514.

Ada Lovelace Institute (2021). Technical methods for regulatory inspection of algorithmic systems in social media platforms.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6: 52138–52160, 2018.

Alan Turing Institute, Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector

Albanie, S., Shakespeare, H., & Gunter, T. (2017). Unknowable manipulators: Social network curator algorithms. *arXiv preprint arXiv:1701.04895*.

Algo:aware (2018). Raising awareness on algorithms. Procured by the European Commission's Directorate-General for Communications Networks, Content and Technology.

AlgorithmWatch (2020). Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse. *Algorithm Watch*

Amnesty International (2021). The Pegasus Project: https://www.amnesty.org/en/latest/press-release/2021/07/the-pegasus-project/

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New media & society*, 20(3), 973-989.

Ananny, M., & Gillespie, T. (2016). Public platforms: Beyond the cycle of shocks and exceptions. *IPP2016 The Platform Society*.

Andreotta, A. J., Kirkham, N., & Rizzi, M. (2021). AI, big data, and the future of consent. *Ai & Society*, 1-14.

Backstrom, L. (2013). News feed FYI: A window into news feed (p. 6). Menlo Park: Facebook for Business.

Balding, C. (2020) 'Chinese Open Source Data Collection, Big Data, And Private Enterprise Work For State Intelligence and Security: The Case of Shenzhen Zhenhua'. Available at SSRN <https://ssrn.com/abstract=3691999>.

Bank M., Duffy F., Leyendecker V., & Silva M. (2021). The Lobby Network: Big Tech's Web of Influence in the EU. Corporate Europe

Observatory and LobbyControl e.V.

Bernstein, A., de Vreese, C., Helberger, N., Schulz, W., Zweig, K., Baden, C., ... & Zueger, T. (2020). Diversity in news recommendations. *arXiv preprint arXiv:2005.09495*.

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 210-219).

Bogost, I. (2015). The cathedral of computation. *The Atlantic*, 15(01).

Bodo, B., Helberger, N., Irion, K., Zuiderveen Borgesius, F., Moller, J., van de Velde, B., ... & de Vreese, C. (2017). Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents. *Yale JL & Tech.*, 19, 133.

Borgesius, F. J. Z., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles?. Internet Policy Review. *Journal on Internet Regulation*, 5(1).

Borgesius, F. J. Z., Möller, J., Kruikemeier, S., Fathaigh, R. Ó., Irion, K., Dobber, T., ... & De Vreese, C. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, 14(1), 82-96.

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proc. Priv. Enhancing Technol.*, (4), 237-254.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3), 209-227.

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press, USA.

Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation.

Bradshaw, S., Howard, P. N., Kollanyi, B., & Neudert, L. M. (2020). Sourcing and automation of political news and information over social media in the United States, 2016-2018. *Political Communication*, 37(2), 173-193.

Bruns, A. (2019). After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11), 1544-1566

Bruns, A. (2019). It's not the technology, stupid: How the 'Echo Chamber'and 'Filter Bubble'metaphors have failed us. *International Association for Media and Communication Research*.

Bunting, M. (2018). From editorial obligation to procedural

accountability: policy approaches to online content in the era of information intermediaries. *Journal of Cyber Policy*, 3(2), 165-186.

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30–44.

Burri, M. (2015). Contemplating a'Public Service Navigator': In Search of New (and Better) Functioning Public Service Media. *International Journal of Communication*, 9, 1341-1359.

Byung-Chul, H. (2016). *Psicopolitica. Il neoliberismo e le nuove tecniche del potere*. Nottetempo, Milano.

Cabral, L. (2021). Merger policy in digital industries. *Information Economics and Policy*, 54, 100866.

Carr, N. (2010). *The shallows: How the internet is changing the way we think, read and remember.* Atlantic Books Ltd.

Castro, J. C. L. (2016). Social networks as dispositives of neoliberal governmentality. *Journal of Media Critiques*, 2(7), 85–102.

Cobbe, J., & Singh, J. (2019). Regulating Recommending: Motivations, Considerations, and Principles. *Considerations, and Principles* (April 15, 2019).

Cobbe, J. (2020). Algorithmic censorship by social platforms: power and resistance. *Philosophy & Technology*, 1-28.

Cwajg C. M. (2020). Transparency Rules in Online Political Advertising: Mapping Global Law and Policy.

Dalton, R. J., & Klingemann, H. D. (2007). Citizens and political behavior. In *The Oxford handbook of political behavior*.

DECODE, Vercellone, C., Brancaccio, F., Giuliani, A., Puletti, F., Rocchi, G., & Vattimo, P. (2018). *Data-driven disruptive commons-based models* (Doctoral dissertation, CNRS).

Deibert, R.J. (2019). The road to digital unfreedom: three painful truths about social media. *J. Democracy,* 30(1), 25–39.

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, 39(9), 761-770.

DeVito, M. A. (2017). From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital Journalism*, 5(6), 753–773.

Diakopoulos, N. (2014). Algorithmic Accountability. *Digital Journalism*, 3(3), 398–415. doi:10.1080/21670811.2014.9764

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62.

Dobber, T., Ó Fathaigh, R., & Zuiderveen Borgesius, F. (2019). The regulation of online political micro-targeting in Europe. *Internet Policy Review*, 8(4).

EDRi (2021). Beyond Debiasing Regulating AI and its inequalities. EDRi: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16, 18

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33), E4512-E4521.

de Cock Buning, M. (2018). A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation. Publications Office of the European Union.

Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, 1-21.

Finck, M. (2018). Digital co-regulation: designing a supranational legal framework for the platform economy. *European Law Review*

Floridi, L. (2016). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics*, 22(6), 1669–1688.

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology*, 31(1), 1-8.

Floridi, L. (2021a). Translating principles into practices of digital ethics: Five risks of being unethical. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 81-90). Springer, Cham.

Floridi, L. (2021b). The End of an Era: from Self-Regulation to Hard Law for the Digital Industry. *Philosophy & Technology*, 1-4.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020-1).

Fogg, B. J., Lee, E., & Marshall, J. (2002). Interactive technology and persuasion. *The Handbook of Persuasion: Theory and Practice*. Thousand

Oaks: Sage

Fuchs, C., & Sandoval, M. (2015). The political economy of capitalist and alternative social media. In *The Routledge companion to alternative and community media* (pp. 165-175). Routledge

Gal, M. S. (2017). Algorithmic challenges to autonomous choice. *Michigan Telecommunications and Technology Law Review*, 2017.

Gillespie, T. (2014). The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167(2014), 167.

Gillespie, T. (2018), 'Regulation of and by platforms', in J. Burgess, A. Marwick and T. Poell (eds), *The SAGE Handbook of Social Media*, London: SAGE, pp. 254–78.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., ... & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4), 4 1, 29.

Google, 'Effect of disabling third-party cookies on publisher revenue' (2020). https://services.google.com/fh/files/misc/disabling_third-party_cookies_publisher_revenue.pdf> accessed 13 November 2020.

González-Cabañas, J., Cuevas, Á., Cuevas, R., López-Fernández, J., & García, D. (2021, November). Unique on Facebook: formulation and evidence of (nano) targeting individual users with non-PII data. In *Proceedings of the 21st ACM Internet Measurement Conference* (pp. 464-479).

Gorwa, R. (2019). What is platform governance?. *Information, Communication & Society*, 22(6), 854-871.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2053951719897945.

Gorwa, R., & Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 12(2), 225-248.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 534). ACM.

Helberger, N., Pierson, J., Poell, T. (2018). Governing online platforms: from contested to cooperative responsibility. *Inf. Soc.* 34(1), 1–14.

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.

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.

Hildén, J. (2021). The Public Service Approach to Recommender Systems: Filtering to Cultivate. *Television & New Media*, 15274764211020106.

HLEG AI (High-Level Expert Group on Artificial Intelligence). (2019). Ethics guidelines for trustworthy AI.

Hildebrandt, M., & Koops, B. J. (2010). The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review*, 73(3), 428-460.

Hildebrandt, M. (2015). *Smart technologies and the end (s) of law: novel entanglements of law and technology*. Edward Elgar Publishing.

Hildebrandt, M. (2022). The Issue of Proxies and Choice Architectures. Why EU law matters for recommender systems. *Frontiers in Artificial Intelligence*, 73.

Hirsch, D. D. (2010). The law and policy of online privacy: Regulation, self-regulation, or co-regulation. *Seattle UL Rev.*, 34, 439.

Hoffmann, C. P., Lutz, C., Meckel, M., & Ranzini, G. (2015). Diversity by choice: Applying a social cognitive perspective to the role of public service media in the digital age. *International Journal of Communication*, 9(1), 1360–1381.

Hwang, T., (2020) *Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of the Internet*. Farrar Straus & Giroux. Rome

De La Chapelle, B. and L. Porciuncula (2021). We Need to Talk About Data: Framing the Debate Around Free Flow of Data and Data Sovereignty. Internet and Jurisdiction Policy Network.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3819-3828)

Kaye, D. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports

Keymolen, E. (2016). Trust on the line: a philosophycal exploration of trust in the networked era

Kidron, B., Evans, A., Afia, J., Adler, J. R., Bowden-Jones, H., Hackett, L., & Scot, Y. (2018). *Disrupted childhood: The cost of persuasive design*. 5Rights.

Kingaby H., and Kaltheuner F. (2020). 'Ad Break For Europe The Race To Regulate Digital Advertising And Fix Online Spaces'. Mozilla Foundation.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790.

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).

Leerssen, Paddy: Platform research access in Article 31 of the Digital Services Act: Sword without a shield?, VerfBlog, 2021/9/07, https://verfassungsblog.de/power-dsa-dma-14/, DOI: 10.17176/20210907-214355-0.

Lessig, L. (1999). *Code: And other laws of cyberspace*. Basic Books.

Loecherbach, F., Moeller, J., Trilling, D., & van Atteveldt, W. (2020). The unified framework of media diversity: A systematic literature review. *Digital Journalism*, 8(5), 605-642.

Luguri, and Strahilevitz (2021). Committee for the Study of Digital Platforms Privacy and Data Protection Subcommittee. Stiegler Report.

Masood, M., Nawaz, M., Malik, K. M., Javed, A., & Irtaza, A. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. *arXiv preprint arXiv:2103.00484*.

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *Ai & Society*, 35(4), 957-967.

Moen, G. M., Ravna, A. K., & Myrstad, F. (2018). *Deceived by design*. Forbrukerrådet.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, 1-28.

Morozov, E. (2017). "Moral panic over fake news hides the real enemy – the digital giants", The Guardian, Jan 08. https://www.theguardian.com/commentisfree/2017/jan/08/blaming-fake-news-not-the-answerdemocracy-crisis

Mosseri, A. 2018. News feed FYI: Bringing people closer together. https://www.facebook.com/business/news/news-feed-fyi-bringingpeople-closer-together

Nahmias, Y., & Perel, M. (2021). The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. *Harv. J. on Legis.*, 58, 145.

Napoli, P. M. (1999). Deconstructing the diversity principle. *Journal of communication*, 49(4), 7-34.

Napoli, P., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*.

Napoli, P. M. (2019). User data as public resource: Implications for social media regulation. *Policy & Internet*, 11(4), 439-459.

Nash, V. (2019). Internet Regulation and the Online Harms White Paper: Stakeholder Workshop Summary. Available at SSRN 3412790.

Neumann N., Tucker C. E., and Whitfield T., 'Frontiers: How effective is third-party consumer profiling? Evidence from field studies' (2019) 38(6) *Marketing Science,* 918.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., & Nielsen, R. K. (2016). Digital news report 2016. Reuters Institute for the Study of Journalism.

Newman, N. (2020). Journalism, media, and technology trends and predictions. Reuters Institute for the Study of Journalism.

Pariser, E. (2011). *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin, Westminster.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard: Harvard University Press.

Rahman, K. S. (2018). Regulating informational infrastructure: Internet platforms as the new public utilities. *Georgetown Law and Technology Review*, 2, 2.

Reviglio, U. (2019). Serendipity as an emerging design principle of the infosphere: challenges and opportunities. *Ethics and Information Technology*, 21(2), 151-166.

Reviglio, U., & Agosti, C. (2020). Thinking outside the Black-box: The case for "algorithmic sovereignty" in social media. *Social Media+ Society*, 6(2), 2056305120915613.

Reviglio, U. (2022). The untamed and discreet role of data brokers in surveillance capitalism: a transnational and interdisciplinary overview.

Internet Policy Review, 11(3).

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender systems handbook* (pp. 1-34). Springer, Boston, MA.

Romer, P. (2021, May 17). Taxing Digital Advertising. adtax.paulromer.net.

Scasserra, S., & Elebi, C. M. (2021). Digital colonialism Analysis of Europe's trade agenda. *Policy*.

Scott Morton, F., Bouvier, P., Ezrachi, A., Jullien, B., Katz, R., Kimmelman, G., ... & Morgenstern, J. (2019). Committee for the study of digital platforms: Market structure and antitrust subcommittee report. Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business.

Shenkman, C., Thakur, D., & Llansó, E. (2021). Do You See What I See?. Executive Summary. *Policycommons.net*.

Statista Research Department, Nov 1, 2021: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Sunstein, C.R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, Princeton. 4.

Taddeo, M. (2012). Information warfare: A philosophical perspective. *Philosophy & Technology*, 25(1), 105-120.

Taylor, L., Floridi, L., & Van der Sloot, B. (Eds.). (2016). *Group privacy: New challenges of data technologies* (Vol. 126). Springer.

Thaler, R. H., & Sunstein, C. R. (2019). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.

Thurman, N. (2011). Making 'The Daily Me': Technology, economics and habit in the mainstream assimilation of personalized news. *Journalism*, 12(4), 395–415

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism*, 7(4), 447-469.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., ... & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature*.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1).

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. *Social Issues and Policy Review,* 11(1), 274–302.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report,* 27, 1-107.

Wieringa, M. (2020, January). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 1-18).

World Bank (2021). World Development Report 2021: Data for Better Lives. Washington, DC: World Bank.

Yesilada, M., & Lewandowsky, S. (2021). A systematic review: The YouTube recommender system and pathways to problematic content.

Yeung, K. (2017). 'Hypernudge': big data as a mode of regulation by design. *Inf. Commun. Soc.* 20 (1), 118–136.

Youyu, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans in *Proceedings of the National Academy of Sciences*, 112(4), 1036-1040.

Zarsky, T.Z. (2019). Privacy and manipulation in the digital age. *Theor. Inq. Law* 20(1), 157–188.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology*, 30(1), 75-89.

Box 1 — The Reductionism of Profiling Technologies

Profiling technologies that allow personalization and recommendations create a kind of knowledge that is inherently probabilistic. As a matter of fact, profiling technologies cannot produce or detect a sense of self. Data used are derived only from what is actually observed and are a small subset of possibly observed behaviors. Then, this narrow subset of recorded behaviors must be converted into a digital representation. In this process, information may be lost or misinterpreted. Eventually, four key characteristics of profiling technologies arise:

- → Profiling is highly behaviorist its assumptions. Behaviorism seeks to eliminate theoretical causal mechanisms of human behavior (beliefs, intentions, goals, etc.) and focus instead on what can be observed measured and recorded. Behaviorism actually "confuses the modest proposition that we only have access to what we can observe with the claim that only that which we can observe exists and/or matters" (**Hildebrandt, 2021, p.7**). As such, behaviors most amenable to measurement tend to be recorded whereas what is inside the individual is mostly ignored. Profiling focuses on predicting a very narrow set of possible behaviors, often limited by the context of the application. This can make it seem more powerful and accurate than it really is, especially when predictive performance is evaluated. Most of these systems, for example, tend to capture mostly positive feedback.
- → Profiling is meant to infer users' "preferences", roughly meaning "what people want." This, however, has been justified by the assumption that people always choose what they want, an idea from 20th-century economics called "revealed preferences". Not only this approach is disputble but it can also lead to a variety of unwanted outcomes including clickbait, addiction, or algorithmic manipulation.
- → Profiling becomes ever more persuasive thanks to the emergence of psychographic techniques and "emotional AI". These techniques, however, often rely on a highly contested scientific paradigm that argues that all humans, everywhere, experience the same basic emotions, and express those emotions in the same way. Those emotions include happiness, anger, sadness, disgust, surprise, and fear. This paradigm of universal emotions is insufficiently evidence-based and poorly regarded in the relevant scientific communities.
- → Profiling uses data not only from the individual user, but also from other users. This is clear in the case of RSs using social network data (e.g. collaborative filtering). This lead to question the extent to which recommendations are actually personalized for individuals.

The underlined – somehow inevitable – reductionism is concerning for it can undermine the development of the self, the formation of a healthy and informed public opinion and ultimately serve only the interests of service providers. These limitations need to be acknowledged and eventually overcome.

Box 2 — The Case of Facebook's Algorithm

Facebook has 2.85 billion active users (**Statista, 2021**) and is likely the most pervasive and powerful intermediary on the Internet. Its unique relevance depends also on the fact that regarding personalization, search engines like Google are, to some extent, less problematic because they tend to deliver more one-size-fits-all services. The system of news updating in Facebook – the NewsFeed – has an ever growing central role in the global information flow. On average, it selects 200 posts on 2.000 recommendable posts each day, therefore hiding 90% of the content. It is based on EdgeRank, a complex algorithm which constantly changes its outputs. In a rather unique analysis, DeVito (**2017**) examined the few publications by Facebook that provided information about the operating principles of the algorithm (e.g., press releases, blogs or patents) and was able to identify nine relevant "editorial values": "Friend relationships, explicitly expressed user interests, prior user engagement, implicitly expressed user preferences, post age, platform priorities, page relationships, negatively expressed preferences, and content quality" (**DeVito, 2017, p. 14**). In this context, of particular interest is their relative importance: the ranking criteria of Facebook show only slight overlaps with classical news factors: only the prioritization of new stories and local stories are related to the news factors "novelty" and "proximity" (**Napoli, 2019**). However, the three most important characteristics mentioned are the quality of the relationships (affinity) as well as explicitly and implicitly stated user interests, the latter being identified by the user's previous behavior. The analysis of Facebook's patents in particular showed that, "friend relationships are a guiding value that mediates the application of all the other values" (**DeVito, 2017, p. 14**). This has been emphasized in an update of the news feed in 2018 when Facebook explicitly strengthened the prominence of posts assumed to stimulate discussions and other "meaningful interactions" such as shares and likes in user's networks (**Mosseri, 2018**). Ultimately, the information available on the selection principles of the Facebook algorithm is only an approximation of the actual selection, as the more than 200 selection criteria are well-guarded company secrets and constantly updated and adapted (**De Vito, 2017**).

Box 3 — The Challenges of Media Diversity

As a reaction to the new mostly personalized media environment, media scholars and policy-makers discussed how to preserve and cultivate media diversity. Generally, this is achieved when users autonomously enjoy a diverse media diet. Yet, diversity is not a simple clear-cut concept. Napoli (**1999**) provided a useful taxonomy, dividing diversity into three categories: source, content, and exposure diversity. Source diversity refers to both the plurality of media sources, their ownership as well as the diversity of the workforce at the media organizations. Content diversity refers to the diversity of programing, the diversity of represented demographic groups and diversity of opinion. Lastly, exposure diversity refers to the extent citizens consume diverse programing. This gained prominence online as the focus of policymakers and academics has partly shifted from spheres of production to those of distribution. And even if the exposure to diversity online is generally more than in traditional media, it does not always end up in an 'experience of diversity' (**Hoffmann et al., 2015**). Cognitive and affective factors that drive Internet users must also be considered. This requires to employ a user-centric perspective and extend beyond the assumption that supply and exposure diversity equals experience of diversity, and that diversity of sources equals diversity of content. As such, scholars like Helberger et al. (2016) suggest that PSM should employ RSs that expose their users to diverse content. But, still, how much exposure to how many different contents and sources can be considered sufficient? Fundamentally, the problem lies on the conceptualization of diversity. Media diversity, in particular, is a rich and complex ideal that can be achieved in many different ways, and its interpretation differs significantly per discipline (**Loecherbach et al., 2020**; **Bernstein et al., 2020**).

While many scholars from different disciplines agree that media diversity is an important value that we should include even in the design of institutions, policies and online services, this value is often reduced to single definitions such as "source diversity" or "hearing the opinion of the other side". There is a need for a more detailed normative conceptualization of this value. Only then, perhaps, it will be possible to translate this complex value into design requirements of information intermediaries and move towards viable solutions that can be implemented. Importantly, there are tensions (trade-offs) in diverse design proposals, for example 'diversity' and 'trustworthiness' on the one hand, and 'non-discrimination' and 'neutrality' on the other. While the former requires algorithmic systems to prioritize certain content, the latter could arguably prohibit the drawing of such distinctions (e.g. conspiracy theories increase diversity). Also, democracy theories have conflicting values (**Helberger, 2019**). Thus, the issue of diversity and its design cannot be 'solved' objectively or definitively, rather, throughout more interdisciplinary experimentations and mutidisciplinary collaborations it will eventually be possible to weight diversity in design and recommendation algorithms and ultimately approximate a more diverse media consumption.

Box 4 — The Rise of Deepfakes

Deepfakes (a portmanteau of "deep learning" and "fake") are synthetic media in which a person in an existing image, video or even audio is replaced with someone else's likeness. In the last few years, they have captured widespread attention for their uses in celebrity pornographic videos, revenge porn and fake news. Due to increased interests from amateurs, commercial companies and even governments, along with video generation easiness and improved video quality, they elicited responses from both industry and government to detect and limit their use as well as from scholars to understand their potential impact (**Fallis, 2020; Vaccari and Chadwick, 2020; Masood et al, 2021**). Still, research is limited and policies are lacking.

Deepfakes mainly threaten to worsen epistemic uncertainty and, as a consequence, to decrese the quality of public opinion's formation. So far, however, most deepfakes currently present on social media platforms may be regarded as harmless, entertaining, or artistic. There are, however, also some examples where deepfakes have been used for revenge porn, hoaxes, for political or non-political influence, and even financial fraud. They undoubtedly have the ability to harm, increase misinformation and epistemic uncertainty while at the same time decrease trust over news, people, politicians, and among foreign governments. Deepfakes have indeed the potential to initiate political tension, conflicts, violence, and even war worldwide. In the worst case scenario, deepfakes may trigger or substantially influence an informational cold warfare for they could sway political elections and initiate or worsen geopolitical conflicts, making people ever more cynical and distrustful and politicians eager to restore "order" and "certainty" through illiberal policies curtailing free speech and other civil rights.

Raising awareness among the population and systems of prevention, detection and control could eventually result sufficient to tame the threats of deepfakes (Masood et al, 2021). As Vaccari and Chadwick (**2020**) argue, it is also possible that the reduction of trust in news on social media resulting from the uncertainty induced by deepfakes may not generate cynicism and alienation, but skepticism. Skepticism can reduce susceptibility to misinformation effects if it prompts people to question the origins of information that may later turn out to be false while at the same time ensuring that accurate information is recognized and valued. While skepticism is no panacea, it is much less problematic for democracy than cynicism and may be a sign, or even a component, of a healthily critical but engaged online civic culture.

Box 5 — The Role of European Union

As the United States remains paralysed by partisanship and the concern to maintain its hegemony, and as the Chinese government aggressively pursues its own distinct approach to digital media based on very different values, the European Union (EU) is increasingly emerging as a global policy entrepreneur on digital issues. As the UN Special Rapporteur on freedom of opinion and expression David Kaye has repeatedly pointed out, from the more active informal and formal regulation of online content to more robust competition and data protection policies, Europe will de facto regulate the global internet (**Kaye, 2019**), in line with the 'brussels effect' which refers to the European global influence on markets setting standards by law, in particular due to the strong Europeans' citizens purchase power (**Bradfort, 2020**).

As policymakers all over the world look to Europe for inspiration, this is a unique opportunity for the EU and its member states to show leadership and demonstrate what truly democratic digital media policies can look like. The EU has indeed set its overarching ambition on a human-centric approach to AI. This endeavior has been captured in the notion of Trustworthy AI, which the High Level Expert Group on Artificial Intelligence characterised in terms of three components – being lawful, ethical and robust – and in line with the core tenets of the European Union: fundamental rights, democracy and the rule of law. The Ethics Guidelines for Trustworthy AI (**2019a**) also constitutes a crucial first step in delineating the type of AI that EU wants and do not want. Another related and relevant policy report (**2019b**) presents a set of policy and investment recommendations on how AI can actually be developed, deployed, fostered and scaled in the EU, all the while maximising its benefits whilst minimising and preventing its risks. For this purpose, they formulated a number of concrete recommendations addressed to the European Institutions and Member States.

There are, of course, several challenges for the leading role of EU in the digital legal arena; firstly, to survive against ever more powerful American and Chinese platforms, EU needs to foster European, cross border platforms and allow European media innovations to the scale and achieve the potential of the Digital Single Market (**Klossa, 2019**). Secondly, the policy process is rather slow; there is a high risk that the innovative regulatory impetus we see in recent draft proposals will be outdated before they are even implemented. Thirdly, there are strong lobbying activities from big tech that require more transparency regulation and public oversight (**Bank et al., 2021**). Fourthly, even the EU can be criticized to be another actor in what is called "digital colonialism", particularly in the African continent (**Scasserra and Elebi, 2021**). All in all, EU has the potential to represent a third innovative pole between US and China but undoubtedly faces serious limits and challenges ahead.

Box 6 — Public Service Media and Personalization Systems

Public Service Media (PSM) is the digital version of Public Service Broadcasting (PSB) and it is defined as a universally available media service. These institutions are especially developed in Europe compared, for example, to the US. Nowadays a majority of PSM in Europe are currently moving in the direction of digital and algorithmic personalization (**Van den Bulck and Moe, 2017**), assuming the role of a "Public Service Navigator", meaning a mechanism for influencing the conditions of access to content, particularly its visibility, discoverability, and usability

(**Burri, 2015**).

Actual specifications for RSs, however, are still to be developed. Due to fears of filter bubbles and echo chambers, the attention of the EBU has been mostly focused on exposure to diversity — at the cost of considering the need for common arenas of discourse, thus the values of universality and publicness, and how broadcasting and interactivity/choice can reinforce each other. There are in fact a number of trade-offs involved in personalization systems that European PSM have to confront with. Many institutions with similar histories and comparable media system frameworks are taking up different positions. Some consider the possibility to reach the value of universality through personalization, while other consider personalization to work against it. Some privilege implicit personalization over explicit personalization, with very different outcomes on PSM goals. Another problem is that PSM cannot readily rely on ready-made RSs, due to the fact that most RSs are commercial and generally promote consumption over other values. They must build systems of their own or modify existing systems. And given that the former is very challenging, it is becoming more common that RSs are provided by external contractors (**Hildén, 2021**).

At the same time, it becomes ever more difficult to compete with commercial RSs which are engagement-driven. The actual risk is to lose ever more audience. Nonetheless, PSM is generally taking the challenge, which is also an opportunity. PSM could provide affordances and tools for users – especially in social media – according to its traditional, democratic values, yet with updated goals and strategies. PSM could thus help to set the standards on a more mature, democratic and not profit-driven information and news consumption. Eventually, it could contribute to solve the main challenges to experience diversity online (see Hoffmann et al., 2015). Yet, this seems unlikely as PSM is composed of several actors that rarely act in concert while mainstream social media would have all the sufficient resources, know-how and reach to develop more effectively RSs driven by PSM values.

Box 7 — The Role of Design

The role of algorithms in affecting public opinion cannot be understood without considering how users interact with the same algorithms, and this usually occurs through design. In fact, not only design choices allow users to explicitly influence their online experience, for example for personalization, but, more broadly, these also influence the online debate as well as the dissemination, production and consumption of information. Obvious examples are the 'like' or the 'share' button. These features are clearly not just features but they also carry symbols and steer behaviors. Design choices and the affordances they allow are paramount to the functioning of the algorithmic public opinion. By clicking and liking users fuel the algorithms, which in their turn generate the information flows fed back to users. Moreover, the clicks and likes fuel the interest and engagement of developers, researchers and advertisers who help to keep the platforms in business. In this context, an ever more influential role is assumed by 'deceptive design' or 'dark patterns'. "Dark patterns" define instances where designers use their knowledge of human behavior (e.g., psychology) and the desires of end-users to implement deceptive functionality that is not in the user's best interest (**Gray et al., 2018**)[20]. U.S. Federal Trade Commissioner Rohit Chopra recently defined dark patterns as "design features used to deceive, steer, or manipulate users into behavior that is profitable for an online service, but often harmful to users or contrary to their intent." These are employed extensively not only during terms of conditions and privacy updates (**Moen et al., 2018**), but also for what concerns personalization and its features. A helpful taxonomy, developed by Christoph Bösch and co-authors (**2016**), identifies classic types of privacy dark patterns including bad defaults (which we propose a framework for identifying above, and one example of which is a choice between "Yes" and "Not Now" rather than "Yes" and "No"), privacy zuckering (i.e., providing users options to adjust their privacy settings that are needlessly complex, granular, or confusing), forced account registration (seeming to require registration to use a service), hidden fees or terms added at the end of a long transaction (how did that wind up in my online shopping cart?), forced account preservation (making it impossible to delete accounts once created), and address book leeching (requesting users' contacts at the time of activation and then spamming users' contacts with email invitations) . Design can also help to nudge users without introducing manipulative measures. Such approach is embraced by the philosopher of information Luciano Floridi (**2016**) who advocates for what he calls 'pro-ethical design' or "tolerant paternalism" that, in short, it is the attempt to modify the level of abstraction of the choice architecture by educating users to make their own critical choices and to assume explicit responsibilities. This approach is mainly grounded on the nudge theory (**Thaler and Sunstein, 2019**) and ultimately based on behavioralist assumptions (*see Box 1 — The Reductionism of Profiling Technologies*) which are debatable whether they are right and even democratic at all (**Hildebrandt, 2021**).

To conclude, in the regulation and governance of algorithms the role of design cannot be ignored. Policy-makers and designers are increasingly aware of this; "design policy" is a new area that looks at the role of design in products and software and then analyzes how design operate in relation to policy and technology. In other words, it is the act of using design to make policies around software and hardware understandable. Design policy indeed recognizes that design affects technology, including how technology looks and feels to consumers, and what consumers can 'see' or know about technology.

Box 8 — The Role of Ethics

In the struggle of policy-makers to catch up with the rapid pace of technological innovation, ethics acquires an ever more essential role of moral evalutation so as to complement and improve digital and AI governance. The increasing demand for reflection and clear policies on the impact of AI on society has produced several initiatives that state principles, values, or tenets to guide the development and adoption of AI, including algorithms. Their promise is to condense complex ethical considerations or requirements into formats accessible to a significant portion of society, including both the developers and users. The inventory of AI ethics guidelines compiled by Algorithm Watch lists over 160, of which 88% having been released after 2016, and is constantly being updated. An obvious risk is unnecessary repetition and overlap, if the various sets of principles are similar, or confusion and ambiguity, if they differ.

Analyzing 36 sets of AI principles Fjeld et al. (**2020**) noted a 'convergence' around 8 key themes shared by the various sets of principles: Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, Human Control of Technology, Professional Responsibility and Promotion of Human Values. While we could see this apparent convergence as a sign that we are moving towards some common ethical foundation, we need to acknowledge that it is far easier for companies and governments to sign up to relatively vague ethical principles than it is for them to change business practices or enforce restrictive laws. If principles committing companies to fairness and non-discrimination had serious legal consequences, there should have been far more heated disagreement about precisely how to define these principles. Concepts such as fairness in machine learning are not only contentious - there are indeed 21 different definitions! - but researchers have proven that different definitions of fairness are in fact mutually exclusive.

There are two major risks arising in this context; on the one hand, 'ethics bluewashing', as the practice of fabricating or exaggerating a company's interest in equitable AI systems that work for everyone" (**Floridi, 2021a, p.3**). Ethics guidelines thus would serve as means to dodge regulation. On the other hand, there is the risk of "ethics bashing" as "a tendency, common amongst social scientists and non-philosophers, to trivialize "ethics" and "moral philosophy" by reducing more capacious forms of moral inquiry to the narrow conventional heuristics or misused corporate language they seek to criticize" (**Bietti, 2020, p.221**). Grappling with the role of philosophy and ethics in tech policy requires moving beyond both ethics washing and ethics bashing and seeing ethics as a mode of inquiry.

Clear shifts in governmental policy which can be directly traced back to preceding and corresponding sets of AI Ethics Principles, however remain few and far between. Yet, as governmental policy-making takes time, it may simply be premature to gauge (or dismiss) their impact. It is true, however, that most attention for all the ethical principles is focused on interventions at the early input stages but very few tools or methods during the middle building and testing phases (**Morley et al., 2019**). Also, these conversations are frequently bound to the academic community and related discourses, making practitioner access to these conversations difficult. Comprehensive ethics education is critical to ensure that future generations of practitioners, designers and engineers take their role as creators of futures seriously. We still face a significant gap in background and understanding between, on the one hand, people from the humanities and social sciences, and, on the other hand, people from the natural and engineering sciences, both within and outside academia.

Box 9 — The Future of Personalization Algorithms

Since algorithms are continuously being developed, it is possible that problems caused by (news) personalization could become even more threatening in the future than is currently the case. A focus on the potential development of personalization algorithms, therefore, becomes essential. If we imagine the future of personalization, we can quite confidently argue that algorithms – especially smart assistants such as Alexa or Siri – will guide us and make ever more sophisticated and reliable decisions for us and about us. It is often imagined that algorithms will wake up us softly with our most preferred music, make the perfect breakfast, as our physician recommended, choose the street we take, the people we meet, the books and news we read, similarly to nowadays, but even suggesting important choices of life – what to study, whether to accept a job offer or whether to marry.

These kind of cognitive outsourcing are provided by AI-driven algorithmic decision-making and decision-guidance. They promise us to enhance our lives preserving our time and energy and nudging us towards healthier behaviors and better decisions. A problematic issue with the future of personalization algorithms is that they can increasingly gratify us to the extent that we may come to accept its pervasive and at times deceptive role as benevolent. This argument has been discussed in terms of "psychological hedonism" (**Gal, 2017**): if personalized systems will become 'pleasure machines' able to predict our choices and simply grant them to us, will we be willing to give up our autonomy? And, if so, under which conditions? Emerging markets and research fields like the Internet of Things (IoT), cognitive and affective computing (cognitive science and psychology) will play a significant role in the future of these algorithms. Clearly, there is a fair amount of unpredictability in communication technology development, preventing precise predictions regarding what future implementations of personalization algorithms will look like. Yet, it is expected the rising of Ambient Intelligence in conjunction with the Internet of Things. This construct offers a vision in which automatic smart online and offline environments and devices interact with each other, taking an unprecedented number of decisions for us and about us to cater to our inferred preferences, representing a new paradigm in the construction of knowledge (**Hildebrandt and Koops, 2010**). Similarly, a related promise is the predictive ability of algorithms, especially of emotion. Several emotion-decoding technologies are being developed nowadays throughout all the potential inputs: from the face through the analysis of facial expression, micro-gestures, micro-movements of muscles and eye tracking to other inputs such as voice tones, body language, keyboard typing recognition etc. These, however, may still suffer from the significant theoretical limitations highlighted (see *Box 1 — The Reductionism of Profiling Technologies*).

Box 10 — The Role of Messagging App

A significant challenge is represented by encrypted messaging apps, particularly for content moderation and disinformation. For example, Whatsapp, Telegram and Signal can be understood as social media insofar as content sharing among small and large groups, public communication, interpersonal connection, and commercial transactions converge in key features of the app. These platforms gained immense popularity in key markets like India, Brazil, and Indonesia. Such social media's shift towards more private, integrated, and encrypted services opens up new challenges not only for content moderation on the app, but also for the regulation of platforms by governments.

Despite encryption, WhatsApp moderates content both at the account and content level (**Gillespie et al., 2020**). At the account level, the company uses machine learning to detect abusive behaviour, disables over two million accounts per month, and scans unencrypted information such as profile pictures, which has been instrumental in detecting child pornography activity within the app. At the content level, accusations of disinformation and mob violence pushed WhatsApp to implement measures to curb the virality of problematic content. Unregulated virality on the app depends on a combination of affordances: encryption, groups of up to 256 people, and the forward function. Most of the groups on the platform are family and friends, local community and neighbourhood groups of less than ten people. In these intimate spaces, one might think that content moderation would not be needed at all. However, in countries like Brazil, Malaysia and India, some WhatsApp groups can be much larger and act as semi-public forums. The sharing of news and the discussion of politics are popular; this communication takes place among strangers, and context collapse may occur. The combination of large groups and users' unlimited ability to forward messages helped information on WhatsApp be easily shared at scale, potentially encouraging the spread of misinformation.