Artificial Intelligence, Content Moderation, and Freedom of Expression

Llansó, E.; van Hoboken, J.; Leerssen, P.; Harambam, J.

Publication date
2020
Document Version
Final published version

TRANSATLANTIC WORKING GROUP

Artificial Intelligence, Content Moderation, and Freedom of Expression†

Emma Llansó, Center for Democracy and Technology[1]
Joris van Hoboken, Institute for Information Law, Vrije Universiteit Brussels[2]
Paddy Leerssen, Institute for Information Law[3]
Jaron Harambam, Institute for Information Law, University of Amsterdam[4]

February 26, 2020

Contents

† One in a series: A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Introduction

As governments, companies, and people around the world grapple with the challenges of hate speech, terrorist propaganda, and disinformation online, "artificial intelligence" (AI) is often proposed as a key tool for identifying and filtering out problematic content. "AI," however, is not a simple solution or a single type of technology; in policy discussions, it has become a shorthand for an ever-changing suite of techniques for automated detection and analysis of content. Various forms of AI and automation are used in ranking and recommendation systems that curate the massive amounts of content available online. The use of these technologies raises significant questions about the influence of AI on our information environment and, ultimately, on our rights to freedom of expression and access to information.

What follows is a compact position paper, a first version of which was written for the Bellagio, Italy, session of the Transatlantic Working Group (TWG), Nov. 12-16, 2019. It discusses the interface of AI/automation and freedom of expression, focusing on two main areas.

Part I focuses on **content moderation** and the use of automated systems for detecting and evaluating content at scale. Part II focuses on **content curation** and questions about the role of recommendation algorithms in amplifying hate speech, violent extremism, and disinformation.

For both content moderation and content curation, the paper explores the use of AI and other forms of automation. In particular, it focuses on their use in the fight against hate speech, violent extremism, and disinformation. On that basis, we reflect on the need for new freedom of expression safeguards tailored to these new automated forms of speech governance.

This paper and its recommendations take into account the transatlantic nature of the TWG and are therefore intended to speak to both U.S. as well as EU legal contexts. The authors are grateful for the discussions and observations produced by participants in the TWG meeting at Bellagio, many of which have been incorporated into this report and the recommendations that follow. The text of the paper remains the sole responsibility of the authors.

Key Recommendations

1. Speech governance debates should not focus exclusively on AI technology as such, but take into account a broader range of automating technologies and processes, including simpler forms of automation and algorithmic systems.

2. Automation in content moderation should not be mandated in law because the state of the art is neither reliable nor effective.

3. Entities that use automation in content moderation should provide greater transparency about their use of these tools and the consequences they have for users' speech, privacy, and access to information.

    a. Priority should be paid to enabling availability of research data while respecting users' privacy and data protection rights, and in particular to resolving the questions of

whether or how data protection regimes in the U.S. and Europe constrain the sharing of data for research.

4. Policy makers should resist simplistic narratives that lay the blame for harmful content online exclusively with "algorithms." Instead, policy makers should recognize the role of users and communities in creating and enabling harmful content. Solutions to policy challenges such as hate speech, terrorist propaganda, and disinformation will necessarily be multifaceted.

5. Regulating ranking algorithms raises free speech risks comparable to outright removal of content and should be treated with comparable safeguards. Content regulation via ranking decisions is particularly problematic due to the lack of transparency in this space and the potential for far-reaching and unintended consequences for online discourse.

6. There is no such thing as a "neutral" recommendation algorithm and policy makers should therefore avoid simplistic mandates for "neutrality" or "non-discrimination."

7. Potentially fruitful measures to address harmful content in recommendation systems include:

    a. Enhancing user agency by providing the option to choose between different approaches to ranking.

    b. Media literacy and awareness programs.

    c. Enhancing public transparency about the design and performance of recommendation systems.

Part I: Automation in Content Moderation

Introduction

Enormous amounts of content are uploaded and circulated on the Internet every day, far outpacing any intermediary's ability to have humans analyze content before it is uploaded.[5] Many companies and governments are turning to automated processes to assist in detection and analysis of problematic content, including disinformation, hate speech, and terrorist propaganda.

"AI in content moderation" is a very broad concept: In one sense, there is very little true "artificial intelligence" in use in content moderation systems.[6] As it is typically used in law and policy conversations, however, AI in content moderation can refer to the use of a variety of automated processes at different phases of content moderation. These processes may be as simple as keyword filters, but they may also rely on machine learning and can include a number of tools and techniques. In this section we will examine a variety of machine learning processes and discuss how they are used in different phases of content moderation.

Content moderation on a site or service with a global user base requires the application of a set of rules or standards for speech that may occur in many formats (such as text, images, video, or audio), on any subject, in potentially any language or dialect, from thousands or millions of cultural contexts. As a recent report from Cambridge Consultants noted, content moderation raises a number of challenges, "such as ambiguity within a certain piece of content, the change of meaning within a post

when context is considered and the potential for bias in the moderator. AI moderation has to contend with the same set of contextual challenges, along with a number of AI-specific technological challenges."[7]

In content moderation, automation can be used in the related phases of proactive **detection** of potentially problematic content and the automated **evaluation** and enforcement of a decision to remove, tag/label, demonetize, demote, or prioritize content. Machine learning tools can also be used to automate the **generation** of content and accounts. Understanding these capabilities for creating new material can also aid in the detection of it, including tools for the evaluation of manipulated imagery, such as "deepfakes."[8]

Key concepts in machine learning

Supervised learning involves training a model based on a labeled or annotated training dataset. For example, a research team that is building a tool to identify hate speech on a social media platform could label a corpus of posts as "hate speech" and "not hate speech" (or as "racist," "homophobic," "anti-immigrant," and "not hate speech," or any other set of labels). These labels assist the machine learning model in identifying features of the data that help to distinguish the various categories of posts from one another. Supervised learning models can be effective at developing classifiers to distinguish or categorize different inputs, but they can be resource-intensive to train, as they require substantial quantities of hand-labeled inputs. Further, as discussed below, the processes of generating the training dataset and having one or more people label it can introduce biases and errors into the model.

Unsupervised learning involves training a model based on an unlabeled dataset; the model learns to identify underlying patterns and features within the data. Unsupervised learning can, for example, be used to develop a corpus of word pairs that often occur together in a text. These pairs can then serve as the "labeled" training data for a tool like Word2Vec that assesses the relationships among words called "word embeddings." [9] Word embeddings can be a powerful way to automatically parse text, but such tools can also inadvertently "learn" associations in the underlying text that reinforce cultural biases.[10]

There are a number of different types of algorithms that are frequently used in machine learning. For instance, supervised learning often uses linear regression or decision tree algorithms, whereas unsupervised learning uses K-means clustering or Apriori association rule learning.

Discriminative algorithms are those that apply a classifier to determine whether an input should be labeled as fitting into a particular category (e.g., spam or not spam). *Generative algorithms*, by contrast, start from a particular label and predict what features or characteristics the "input" should have.

Mathematical *models* are what encode the results of the machine learning done on the dataset. A *neural network* is a layered machine learning model akin to the connections among neurons in the human brain. *Convolutional neural networks* (CNN) are a type of neural network commonly used to evaluate image data.[11] *Recurrent neural networks* (RNN) are neural networks that incorporate outputs from the system into subsequent runs and are thus better able to process sequences of information.[12]

Capabilities and limitations of automated content analysis

Analysis of text

Automated blocking and removal of online text predates sophisticated machine learning techniques: some of the earliest automated content moderation systems relied on keyword filtering to block posts or access to websites that included certain words and phrases. Keyword filtering is notoriously overbroad and underinclusive, blocking words regardless of their context or meaning and failing to filter content not specified on the list of prohibited terms.

Natural language processing (NLP) is a field of study that seeks to enable computers to parse text in a more comprehensive way, closer to the way that a human would understand the text.[13] NLP tools can be trained to predict whether a text is expressing a positive or negative emotion (sentiment analysis) and to classify it as belonging or not belonging to some category (such as the hate speech classifier described above). NLP tools are often trained on text that has been stripped of features such as URLs, usernames, and multilingual communication, but some researchers are beginning to experiment with incorporating information from emojis into sentiment-analysis tools.[14]

A well-known example of an NLP tool is Google/Jigsaw's Perspective API, an open-source toolkit that allows website operators, researchers, and others to use Perspective's machine learning models to evaluate the "toxicity" of a post or comment.[15] Perspective provides a good illustration of both the capabilities and limitations of a sophisticated NLP tool. It has been used for a variety of applications,[16] including as a tool that comment moderation systems use to warn users that they may be posting a "toxic" comment and to give them the opportunity to revise their comment.[17] But Perspective, and the concept of evaluating "toxicity" of comments, is far from perfect; soon after the Perspective API was launched, researchers began exploring ways to "deceive" the tool and express negativity that slipped under the radar,[18] and researchers continue to identify bias in the tool, such as misclassification that disproportionately affects different racial groups.[19] The Conversation AI research team behind Perspective cautions, "*We do not recommend using the API as a tool for automated moderation*: the models make too many errors."[20]

Machine learning models can also be used in the generation of text. Earlier this year, the OpenAI research team announced the public release of their GPT-2 language model, a predictive text tool that was trained on a dataset of eight million web pages.[21] The text generated by the GPT-2 model can be fairly complex and, according to the researchers, regularly outperforms the previous state-of-the-art models for text generation.[22] The research team made a widely publicized decision to release only a smaller, less capable model for public scrutiny and use, "[d]ue to concerns about large language models being used to generate deceptive, biased, or abusive language at scale."[23] Six months later, it released a larger version of the model along with a paper discussing the lessons it had learned through this "staged release" strategy which has allowed OpenAI and other researchers to more fully consider the implications and potential malicious uses of this technology.[24] (Meanwhile, two graduate student researchers at Brown University attempted to replicate OpenAI's full GPT-2 model and released their version publicly in August 2019.)[25]

Analysis of images

Automated image-detection and -identification tools can range from fairly simple systems designed to detect previously identified content to more complex tools designed to discover features of novel content. *Hash values* are unique numerical values that are generated by running a specific algorithm on a file; the *hash function* calculates the numerical value based on characteristics of the file, and can be thought of as generating a specific "digital fingerprint" for that file. A system can run the same hash function on novel/newly uploaded content and detect whether the novel content matches the hash value of previously identified content.

For simple hashing, the characteristics that the function evaluates can include things like dimensions of the image and specific color values of pixels; changing any of these characteristics, however, can completely change the hash of the altered file, which makes simpler hashing functions easy to circumvent if the goal is to evade detection.[26] A more sophisticated approach, *perceptual hashing*,[27] can be more resilient against circumvention by calculating hashes based on relationships among pixels and accounting for minor variations in the resulting hash.[28] Microsoft's PhotoDNA tool, for example, converts an image to black and white, resizes it to a standard size, divides the image into a grid, and calculates the hash based on the intensity gradient of each black-and-white square.[29] These transformations of the image, as part of the hash function, help to counteract minute changes to the image that could stymie simpler hash functions.

Other approaches to image analysis, including machine learning methods and techniques from the field of *computer vision*, work to detect the presence of specific elements or features in an image, such as symbols or logos, weapons, or nudity.[30] Tools that are designed to detect pre-identified images, such as a specific symbol or logo, need to be able to identify variations of that symbol in different lighting conditions, resolutions, and rotation/skew. *Optical character recognition* tools can identify text in an image and convert it into machine-readable formats, which is an essential step to using natural language processing to evaluate the meaning of text.

Tools can also be designed to classify whether an image contains a feature such as nudity. One approach to detecting nudity in an image has been to analyze the proportion of pixels in an image that fall into a specific color range that has been pre-identified as representing skin color. This kind of tool is vulnerable to misclassification of underrepresented skin tones and of objects or scenes with the same color palette as the training data (for example, deserts[31]). More involved machine learning tools will use skin-tone detection as a component of their analysis, along with other image-parsing processes to detect, for example, the presence of faces[32] and distribution of body parts. Such tools then use this underlying information about the components of an image to generate a classifier to identify likely nudity or sexual activity. Even more complex tools, however, can erroneously classify content to a significant degree when they are used on mass-scale, highly variable datasets like social media postings, as Tumblr discovered when it implemented its ban on nudity and sexual content, which yielded a wide variety of false positive results.[33]

Other deep-learning methods enable techniques such as scene understanding, which not only identifies the discrete features/likely objects in an image but analyzes them in the context of their relationship with the other objects in the image.[34]

The field of image-generation is also rapidly developing. For example, *Generative Adversarial Networks* (GAN) use generative algorithms to create new data to test and refine a machine learning classifier. GANs can be useful for training a machine learning model in how to detect a manipulated image or video.[35] Generative algorithms can also be used to deceive machine learning tools by manipulating images so that they look essentially unchanged to the human eye but display mathematical features that the classifier will understand to be something else entirely.[36] Researchers have demonstrated effective adversarial techniques even against black-box networks, i.e., networks for which the attackers have no specific knowledge of the model or the training data that generated it.[37]

Image generation also notably includes the area of "deepfakes," composite videos and images created on the basis of real footage that portray fictional statements and actions. A common form of deepfake is face-swapping, where the face of one individual is superimposed on the body of another. For example, in a video aimed at warning people not to believe everything they see on the Internet, filmmaker Jordan Peele created a deepfake video where his words appear to be spoken by President Obama.[38] To accomplish this, an *autoencoder* is used to analyze a large volume of images of a person to create a detailed mathematical map of the features of an individual's face (encoding) and to develop a process for turning these features back into the image of the individual's face (decoding). Once the autoencoders are trained, the encoded data of one person's face can be translated back into an image by the decoding process for another person's face, essentially preserving the features of the first person but placing them in the context of the second person's face.[39] A GAN can then be used to evaluate and iterate on the produced images and video and develop a more refined outcome.[40] Deepfakes can threaten rights to privacy and dignity, as in the case of involuntary pornography,[41] and may be used in disinformation campaigns. However, experts caution that the expense and time required in creating deepfakes will likely limit their use in disinformation campaigns, especially when it is easier to create misleading or recontextualized information to achieve similar results.[42]

There are a variety of other forms of automated content analysis that can be relevant in content moderation, including video and audio analysis techniques and efforts to detect bot networks/accounts.[43]

Technical limitations of automation in content moderation

Different technical approaches to automated detection and analysis of user-generated content will have limitations specific to their design – for example, a tool designed to detect "toxic" comments in one language will have difficulty parsing multilingual text.[44] This section summarizes major technical and design limitations to automated content detection and analysis; Section 1.3 examines the broader limitations of these tools and the risks for freedom of expression when they are incorporated into content moderation systems at mass scale.

The importance of context: Whether a particular post amounts to a violation of law or content policy often depends on context that the machine learning tool does not use in its analysis. Some context could be incorporated into a machine learning tool's analysis, such as the identity of the speaker or the relationship between sender and receiver of a message, but these come with significant tradeoffs for privacy. Other context, such as historical, political, and cultural context, are much more difficult for a tool to be trained to detect.

Lack of representative, well-annotated datasets to use for training: Machine learning tools develop their ability to identify and distinguish different kinds of content based on the datasets they are trained on. Many tools are trained on labeled datasets that are already publicly available; if these datasets do not include examples of speech in different languages and from different groups or communities, the resulting tools will not be equipped to parse these groups' communication.

Annotation for supervised learning can introduce bias: The process of labeling a dataset for supervised learning typically requires the involvement of multiple human beings to evaluate examples and select the appropriate label, or to evaluate an automatically applied label. *Intercoder reliability* is an important measure of how consistently different humans involved in labeling a dataset perform this task. Low intercoder reliability means that the humans applying the label do not agree among themselves what content merits the label of, for example, "hate speech" or "spam."

Need for flexible, dynamic models: Human communication patterns can change quickly, and speakers who are blocked by automated filters often have extra incentive to figure out how to circumvent the filter. Static machine learning models will quickly become outdated and unable to correctly classify users' communications.

Domain specificity: Natural language processing tools perform best in environments that closely match the data they were trained on. It is difficult to develop tools that work well across a variety of sites, languages, cultures, interest groups, and subject matter.

Significant risks of bias against underrepresented speakers: Applying a tool to a domain or group of speakers who do not closely match the groups represented in the training data can lead to erroneous classifications that disproportionately affect underrepresented groups.[45]

Resource limitations in energy, data, and processing power: Geoffrey Hinton, "one of the forefathers of modern deep learning," argues that we may have hit the limits of what existing machine learning techniques can do. "Hinton points out that CNNs [convolutional neural networks] are highly inefficient at learning features: neural networks require huge amounts of memory and computing power, and massive amounts of data, and still struggle with translational and rotational changes of objects."[46] Wide-scale application of ultra-sophisticated machine learning models for content analysis may be too resource- and energy-intensive to be sustainable.

Analysis: Freedom of expression threats and safeguards

Beyond the technical limitations of any particular tool, the use of automation in content moderation systems raises distinct challenges for freedom of expression and access to information online. There is a growing body of literature on the human rights implications of the use of automation by online services: The UN Special Rapporteur on freedom of expression, David Kaye, has offered analysis, conclusions, and recommendations on artificial intelligence in a number of recent reports.[47] The Council of Europe has provided several reports, studies, and recommendations that touch on the topic and is in the process of finalizing a new recommendation on the human rights impacts of algorithmic systems.[48]

It is important to note that the application of artificial intelligence in the online media environment can have both positive and negative implications for individuals' right to freedom of expression. First,

AI is at the core of the services that are central to effectuating people's rights to express themselves and to access information. Search engines, social media, and other internet services deploy various complex and adaptive information-processing technologies at the core of their operations. Without these technologies, these services, and the value they provide to people in expressing themselves and accessing information, would simply not be possible. This is not to say that relevant services are always doing a good job in supporting freedom of expression. There are significant concerns about how well current services serve our democracies and respect people's right to freedom of expression. Even so, the deployment of these advanced data-processing operations will remain central to our media and communications ecosystem.

Second, algorithmic systems have become a necessary tool to defend freedom of expression and the values underlying it. AI not only powers complex service operations, it is increasingly necessary to create the conditions for a robust and vibrant democratic exchange on online platforms. To make a simple analogy: Without AI, our media would increasingly feel like an email inbox without a spam filter.

Third, such filtering systems, including content moderation and recommendation systems (see Section II), raise their own set of freedom of expression concerns. Their application can raise issues of bias and discrimination, private due process, and surveillance issues that our current legal frameworks have not fully addressed. When combined with regulatory pressure on platforms to tackle issues such as disinformation, terrorism content, and hate speech, the application of these tools raises privatized censorship concerns, as well as questions about prior restraint, and due process.

AI tools and risks to freedom of expression

There are a number of recognized issues with the application of algorithmic systems and automation for these purposes:

False positives and false negatives: The use of algorithmic systems for detecting particular types of speech and activity will always have so-called false positives (something is wrongly classified as objectionable) and negatives (the automated tool misses something that should have been classified as objectionable). From a freedom of expression perspective, the implications of false positives and negatives depend on the goals of the tool that is used. If the tools are used to identify and demote or remove content, or single it out along with the relevant content creators for further scrutiny, false positives risk significant burdens on individuals' right to freedom of expression. False negatives, on the other hand, can result in a failure to address hate speech, harassment, and other objectionable content that may create a chilling effect on some individuals' and groups' willingness to participate online.

Potential bias and algorithmic discrimination: Algorithmic systems have the potential to perform badly on data related to underrepresented groups, including racial and ethnic minorities, non-dominant languages, and/or political leanings. This is due both to the lack of data and to the possibility of biased training datasets: If data are influenced by real-world biases and inequalities, then the models trained on these data may come to reflect or amplify these inequalities. This can result in

serious risks to freedom of expression for communities and individuals, potentially including illegitimate silencing of their expression and failure to address harms to their communities.

Large-scale processing of user data and profiling: Algorithmic systems will typically rely on the large-scale processing of user data to develop and apply the tools. These systems may also involve the additional profiling of users in view of the risk that such users engage in activity that may warrant additional scrutiny and/or risks from a content moderation perspective. In this way, the growing reliance on algorithmic systems further encourages the collection and processing of personal data, which pose additional risks to the rights to privacy and freedom of expression.

Presumption of the appropriateness of prior censorship: Automatically pre-judging content and prohibiting it from being posted is the very definition of prior restraint/censorship. While pre-screening content to limit the spread of malware, child abuse material, and spam has been broadly accepted as a positive use of automation, we must be cautious about applying that logic to other types of speech.

Inadequate oversight and lack of due process: Algorithmic systems may be applied as a quick fix for the complex task of judging whether particular content or activity warrants restrictive actions (which can range from flagging, demoting, demonetizing or removing it, to actions taken against account holders). Without proper complaint, review, and appeal procedures as well as oversight, these actions may violate freedom of expression rights, due, for instance, to the issue of false positives and negatives identified above. In particular, this raises due process issues. And given the difficulty of explaining and documenting complex machine learning systems, it may be even more difficult to create transparency and monitoring here than for other types of content moderation. Fundamentally, there is a tension between the evaluation of AI tools from a statistical perspective (how well are they performing overall) and their evaluation on a case-by-case basis, which is the predominant mode of evaluation from a fundamental rights perspective.

Need for redress and accountability: Automation in content moderation challenges preexisting structures and frameworks for making determinations about speech, given the enormous scale of speech that is being evaluated. What do remedy – and accountability – look like at scale?

The role of platforms in communications governance

Online platforms facilitate and shape the power of the speech of others.[49] They can act (or perhaps better, be asked or forced to act) as control points in tackling the proliferation of illegal and different forms of harmful speech and activity. What their role should be, and by what standards they should be judged, remains a topic of intense debate.

In part because of their critical role in facilitating speech, online platforms are also at the center of discussions about objectionable content. Not surprisingly, online platforms are now central sites for the development of automated content moderation systems and solutions. To develop these systems, larger platforms such as Google and Facebook can leverage world-leading expertise in machine learning and unrivaled financial resources in combination with large datasets and internet-scale user activity. The role of these private sector actors in developing content moderation tools may result in

a lack of transparency and accountability with respect to their functions and effects, including on freedom of expression. In addition, the development and implementation of these technologies will be informed by specific commercial motivations that are connected to the platforms' business models. Any freedom of expression safeguards with respect to the use of algorithmic systems in content moderation will have to address the central role of platforms in the development and application of these tools in real-world settings.

Recommendations

The human rights framework

The implications of artificial intelligence in media and communications for human rights have already been explored. As the UN special rapporteur on freedom of expression concluded in his thematic report to the General Assembly,

> AI tools, like all technologies, must be designed, developed and deployed so as to be consistent with the obligations of States and the responsibilities of private actors under international human rights law. Human rights law imposes on States both negative obligations to refrain from implementing measures that interfere with the exercise of freedom of opinion and expression and positive obligations to promote the rights to freedom of opinion and expression and to protect their exercise.[50]

Further development and deployment of algorithmic systems will continue to raise new issues under the existing human rights framework that should be addressed. One important aspect to take into account is the positive obligation for governments to encourage pluralism and diversity. The online information ecosystem raises a variety of questions with regard to pluralism and diversity: On the one hand, many speakers and perspectives have been able to attain a broader audience than ever before. But on the other hand, the dominance of a few major online advertising providers has radically changed the business model, and fiscal stability, of news and media publishers around the world. Moreover, the application of algorithmic systems for content moderation raises distinct questions of pluralism, as these systems impose a particular set of content restrictions at potentially enormous scale. Policy makers should evaluate the risk to pluralism and diversity that may come from automated content moderation on massive platforms. Policy makers and platform operators alike should consider specific safeguards designed to achieve pluralism and diversity. Such safeguards could range from clear, minimally restrictive default settings combined with greater user controls or Application Programming Interfaces (APIs) that enable users to implement a third-party's content moderation rules on the content they view on a platform.

Regulation of platforms vs. other forms of regulation

Increasingly, regulation is targeting platforms, due to their central role and significant power in the media and communications landscape. Many of the self- and co-regulatory initiatives in this area are also coming from platforms. This can leave other (and existing) forms of regulation to address the fundamental rights implications of algorithmic systems for content moderation unexplored. In Europe, the General Data Protection Regulation (GDPR), which is broadly applicable to the processing of user data in the online environment, provides an important baseline for the protection

of fundamental rights in data-driven power dynamics between platforms and their users. It sets limits to profiling activities, imposes transparency and accountability requirements, and grants relevant rights to individuals. Non-discrimination law has an important role to play in addressing bias and discriminatory impacts of algorithmic systems. Antitrust law should help to prevent undue concentrations of power over media and communications. It's crucial to acknowledge the importance of a broad set of regulatory frameworks that provide important baseline conditions for the effective exercise of freedom of expression online.

Intermediary liability

Intermediary liability frameworks, discussed in a separate TWG discussion paper,[51] have always had a strong link to the state-of-the-art of technologies used to address illegal third-party content and activity. In the early days of the commercial web, both the U.S. and the EU legal systems considered the idea of imposing general filtering mandates on intermediaries that required them to screen out illegal content, but ultimately rejected such mandates due to the significant burden on freedom of expression imposed by overbroad, imprecise filtering. Today, however, legislators are increasingly revisiting the presumption against filtering,[52] based in part on assumptions that filtering technologies have become more sophisticated. Proposed requirements for intermediaries to use filters are based on the rationale that, if there are suitable technologies available to address illegal content and the harms that may result from it, a refusal or unwillingness to use filters could be considered a form of negligence on the part of the intermediary. However, given the persistent risk to free expression posed by the automated detection and evaluation of speech, the deployment of filtering systems without regard for their harmful effects could also be considered a form of negligence. Intermediary liability laws should neither mandate, nor condition liability protection on, the use of filters.

Mandatory and voluntary use of automated systems

The mandatory use of particular forms of automation in content regulation raises prior restraint issues from a freedom of expression perspective. In fact, the mandatory application of such algorithmic systems, the automatic evaluation of content, and the subsequent restrictions on such content from being posted and disseminated fits the very definition of prior restraint, which is generally prohibited under freedom of expression doctrines. In view of this, lawmakers should refrain from imposing mandatory obligations on service providers to impose restrictions on speech through automated systems. Preference should be given to the voluntary deployment of such tools, in combination with safeguards for the fundamental rights of internet users, in view of the issues of false positives, due process, discrimination, and surveillance outlined above.

Systemic responsibilities for platforms and risk assessments

As noted in the paper on intermediary liability, some recent European laws and proposals move away from penalizing platforms for individual incorrect decisions about specific user expression, and instead seek to regulate platforms' overall content management operations and create new forms of administrative oversight with respect to these frameworks. The development of standards for transparency and accountability of content moderation practices is central to this.[53] One accountability mechanism that should be further explored is the use of human rights due diligence processes and other risk assessment methodologies. Risk assessment may be deployed with respect

to particular products or services that a platform develops as well as with particular known threats such as hate speech, terrorism content, and disinformation. Formal risk assessment procedures provide an opportunity for in-depth consideration of the potential impact on fundamental rights that a product or policy poses as well as the various measures that are and may be taken to address them.[54] It is crucial that freedom of expression and associated human rights, and the risks to the rights and freedoms of individuals, are included in these risk assessments. This prevents such risk assessments from being only focused on one particular set of harms and ignoring the potential harmful impacts of mitigating measures on freedom of expression and other fundamental rights.

Complaint and redress mechanisms

The use of automation in content moderation can give rise to additional problems of lack of due process for particular user content and activity. As a result, the development of new complaint procedures and new dispute resolution mechanisms has become a central part of the discussion about online content moderation and freedom of expression. Without proper complaint mechanisms, the actions that platforms take with respect to content and activity lacks accountability and endangers people's ability to effectively exercise their right to freedom of expression. The need for human oversight over automated decision-making deserves a particular emphasis in this context. It is paramount that the results of automated decision-making in individual cases can be scrutinized by humans and that wrong decisions are remedied, both at the individual decision level and through review of the systems that produced the error. Human oversight is necessary in all stages of the development and deployment of algorithmic systems. The possibility of human oversight when there are complaints about the function and impacts of algorithms for content moderation can provide a crucial safety net for the rights and freedoms of affected users.

Freedom of expression by design

"Regulation by design" has become an important way to safeguard fundamental rights. In the case of data privacy, regulation-by-design approaches have matured over the last two decades. These approaches have become an important foundation in privacy law and policy, underpinning data minimization strategies, information accountability, and privacy-friendly approaches to human computer interaction design. In the case of freedom of expression, regulation by design is still in its infancy. The development and deployment of AI and algorithmic systems more generally in the area of content moderation should draw lessons from the regulation-by-design literature developed in this area. By better incorporating freedom of expression and associated human rights concerns into the design and deployment of relevant tools and practices, the effective protection of freedom of expression can be better realized. To achieve this, relevant experts involved in the development and deployment of technologies for content moderation should be supported in developing a more robust field of "freedom of expression by design" approaches. In addition to transparency, accountability, and fairness, such approaches could extend to the creation of relevant datasets and labelling practices, as well as the particular ways in which algorithmic systems are being deployed.

Part II: Content Curation through Recommendation Algorithms

Introduction

Automation is used not only to moderate content, but also to recommend content to potential viewers. Important recommendation systems include content "feeds," such as Facebook's Newsfeed and YouTube's Recommended Videos, as well as search engines such as Google Search. These algorithmic tools play a central role in determining what content is seen online, and what remains hidden.

Content recommendations serve an important need in online media: helping people find relevant content. Based on certain personal and contextual information, these algorithmic systems search through an abundance of content to provide personalized selections of items that are (predicted to be) relevant for the user.

This is by no means a neutral process. Recommendation systems are designed and deployed by specific people in a specific context and for specific purposes. Accordingly, their content selections are dependent on a wide range of factors, including commercial and in some cases political considerations. A number of studies suggest that recommendation systems may funnel people toward disinformation, hate speech and violent extremism.[55] The following sections will elaborate on the functioning of recommendation systems, evaluate existing efforts to remedy their potentially radicalizing qualities, and offer alternative frameworks to mitigate their amplifying dynamics while minimizing the impact on freedom of expression.

Platforms also use (machine learning) algorithms to serve advertisements and other promoted content. This technology allows ad buyers, using large sets of personal data, to microtarget their advertisements toward highly specific audiences in ways that raise concerns about privacy, accountability, manipulation, discrimination, and bias.[56] Yet, despite both recommendation systems and advertising systems using algorithms to personalize their offerings, these types of content distribution are fundamentally different – in terms of substance, sources, functions, relevant harms, governance systems, applicable law, and so forth. Therefore, advertising and sponsored content exceeds the scope of this contribution, which focuses instead on content recommendations for *organic* content, i.e., content that is disseminated without payment to the platform.

Explaining recommendation systems and their deployment

Recommendation systems are automated tools that present ("curate") a selection of content ("recommendations") from an abundance of content. These personalized selections are the result of two main processes. The first is the collection of information about an individual user or browser and the development of a profile based on that information. This involves analyzing data from the user and others like them (website visits, articles read, social media behavior such as clicks and likes). This data is sometimes received from third-party data brokers who sell the data to the entity creating the profile. The second process involves matching the user with the larger pool of content from which the recommendations are drawn. The software system computes a similarity score between the user profile and the characteristics of each individual content item.[57] Recommendation systems are presented as tools to help users find more "relevant" content, but they can also serve other goals.

Most importantly, they help content providers and platforms generate advertising revenue by increasing user engagement.[58]

Proprietary concerns related to this process may make it hard to pinpoint what recommendation systems are actually composed of, and how they function. In turn, this creates serious complications and limits to solutions aimed at transparency.[59] Platforms may be unwilling to lay bare the workings of their recommendation algorithms for proprietary reasons. Even without this impediment recommendation systems are complex systems to grasp, explain, and hold accountable,[60] and some have argued that this narrative can also be a deflection strategy employed by platforms to avoid scrutiny of their systems.[61]

The way recommendation systems work differs between platforms. While search engines like Google deliver filtered selections in response to user queries, social media platforms like Facebook and Twitter give personalized feeds of content independent from any explicit user input. These active and passive recommendation systems both filter out and prioritize content according to specific algorithmic procedures that are optimized for personal "relevance." However, what relevance *means* is not always explained in detail to the public, allegedly because these algorithmic relevancy formulas are well-kept corporate secrets.[62] Users might get *some* explanation as to why they receive a certain post, but the level of detail varies between platforms and services. In some cases, relatively detailed information is published and can offer helpful guidance for sophisticated users to optimize their content for findability.[63] But other services operate largely as "black boxes," certainly when it comes to explaining these complex systems in a meaningful way to the average internet user. Basic disclaimers and notices ("you see this post because you like politics") may not be not very informative about the precise selection criteria at work. What their secret recipe involves, precisely, is often unclear. As discussed further below, a variety of interests and considerations play a role here: recommendation systems are designed for audience engagement, but increasingly also as a vehicle for content moderation or curation.[64]

Recommendation systems are widely used today in various contexts, ranging from commerce (Amazon), travel (Booking.com), music (Spotify), and video on demand (Netflix) to, most relevant for a discussion of disinformation and hate speech, social media platforms (Facebook, Twitter, and YouTube). These social media platforms rely on advertising revenue and their business model centers on engagement, or keeping users "hooked." Controversial, provocative, and extreme content can drive engagement, so their algorithms learn to prioritize such content when it is popular with users.[65]

It should be noted that recommendation systems do not work in isolation, but interact with other human and non-human actors in complex ways. First, platform recommendation systems depend on the content uploaded and shared by users, including content which may be contentious, harmful, or unlawful. The pool of content from which recommendation systems draw is not a neutral representation of public opinion, but biased toward the interests of its most active contributors.[66] And this content pool can tend toward fringe views, if only because there is a relatively weaker incentive to share content displaying scientific or societal consensus.[67] For example: one will find less content arguing that the earth is round, that vaccinations are safe, or that the Clintons are not pedophiles or that there is no relation between race and IQ. The opposing views, however, are overrepresented. This skews recommendations toward the extreme content that is available, even

with hypothetically neutral algorithms. Indeed, research shows that sophisticated users can exploit "data voids" in the content supply: by targeting obscure terms and topics with little to no existing results, they can attempt to draw attention toward potentially manipulative or harmful content.[68] Second, users also influence recommendation systems by behaviors such as liking, rating, sharing, following, scrolling, reading, and clicking; their algorithms are tuned to take popularity (or virality) into account.[69] Strategic actors can therefore deploy bots which send, (re)tweet, like, and spread contentious content to artificially amplify the popularity of certain topics.[70] In addition, research has shown that "false news" spreads much faster and further than "truthful" news, at least within certain user populations.[71]

All this means that there is no such thing as a completely neutral, objective, or unbiased recommendation system: they necessarily reflect and amplify certain preferences, biases, and intentional distortions introduced by their users. In this regard, content recommendations are not entirely unlike editorial decisions made in traditional media; they involve a judgement of relevance and newsworthiness that is necessarily value-laden.[72] (Of course, editors and recommendation systems also differ in many regards, including that recommendation systems are automated and process third-party content, and as a result are generally less intentional or deliberate about overall outcomes.)

The effects of recommended content are highly unpredictable. Nevertheless, actors with more oversight, resources, and/or knowledge of platform dynamics are generally better able to manipulate the wider information landscape. This threat became visible during the "fake news" controversy around the 2016 U.S. presidential election and UK's Brexit referendum, and the subsequent Cambridge Analytica/Facebook scandal,[73] after which it became clear that certain governmental actors had strategically exploited the underlying dynamics of platforms and intentionally inserted dubious claims and outright propaganda in the media ecosystem for political and/or commercial goals.[74] The effects of such interventions on public opinion and politics are difficult to measure or quantify. But these stories do underscore that amplification of harmful content through recommendation systems is not necessarily fully intentional or expected on the part of the platform; it may occur without the platform's full knowledge, intent and control, since these systems operate in complex and dynamic networks of multiple invisible actors and incentives.[75] Of course, deploying these systems without proper research into their potentially harmful effects could still raise a charge of negligence or recklessness.

The complex role of recommendation systems in online hate speech and disinformation

Research on the role of recommendation systems in the circulation of disinformation and hate speech online has surged in recent years. However, studying these systems has proven difficult because of the complexity and magnitude of the current information ecosystem and its ever-changing recommendation algorithms, and the limited cooperation from social media platforms with research communities. Despite these challenges, more researchers within academia and beyond are developing ways to measure and understand how recommendation algorithms contribute to the spread of (dis)information.

Wittingly or not, platforms may actively contribute to the amplification of incendiary, controversial and divisive (dis)information as it directly aligns with the commercial and technological

infrastructures of their recommendation systems that are optimized for user engagement. However, blaming recommendation systems alone ignores the fact that these infrastructures work in conjunction with users' own biased content and behavior, and are furthermore used and strategically exploited by sophisticated actors with more resources and experience than the average user, who can accordingly work the system and gain more political influence.

To give a few examples: Guillaume Chaslot, a former YouTube engineer, developed an algorithmic method to show which recommendations YouTube gives on certain popular topics.[76] He showed that by inserting rather ordinary queries, people increasingly get more and more extreme items recommended, in part due to the content-availability asymmetries noted earlier: asking about "vaccine facts," it takes only a few steps to get to anti-vaxxer conspiracy theories; "global warming" to climate change denialism; "US presidential election" to pro-Trump videos, and entering "the pope" gives suggested videos describing the Catholic leader as "evil," "satanic," or "the anti-Christ."[77] Chaslot's samples are not representative, but nonetheless provide a revealing snapshot that lends empirical support to the theory that YouTube is a "great radicalizer."[78] Since these events, YouTube has revised its algorithms in an attempt to counter extremism.[79]

These findings have been corroborated by a number of other, non-U.S.-focused studies. Kaiser and Rauchfleisch (2017) report on how people watching videos of the populist right-wing party Alternative für Deutschland are recommended by YouTube to watch videos by the vastly more extreme and openly anti-Semitic National Democratic Party of Germany (NPD).[80] Social-media analyst Ray Serrato used computational methods to study the recommendations given by YouTube when searching for "Chemnitz," the East German city where violent anti-immigrant protests erupted in 2018.[81] He shows how ordinary viewers searching for this term were led by YouTube toward more and more extreme videos, while a tightly networked ecology of users and channels was able to amplify the reach and virality of right-wing videos. And investigative journalists from the Dutch news outlets de Correspondent and de Volkskrant undertook a major study into YouTube's "radicalization problem."[82] They gathered massive amounts of data (660,000 videos from 1,500 channels, with 120 million reactions, 15 million recommendations, and 440,000 video transcripts). Their analyses revealed the presence of a tightly knit right-wing reactionary network on YouTube, which lured viewers into a right-wing maze of more and more extreme videos with the help of YouTube's recommendation algorithms. They also showed the radicalization process of certain "heavy commenters" based on content analyses of their comments over a yearlong period. Their efforts to study the radicalizing effects of the recommendation algorithms did not yield conclusive results, which they refrained from publishing.

These studies reveal the difficulty of isolating the influence of recommendation algorithms in the amplification of hate speech and disinformation from the strategic actions of those who use YouTube's platform to convince viewers of their ideology. This intricate entanglement of sociological and technological factors is particularly made clear by the research done by Rebecca Lewis at the Data & Society Research Institute (2018). She identifies a wide, interconnected network of "alternative political influencers" on YouTube who promote alt-right ideologies with specific online branding techniques (testimonials, controversy, platform optimization). Having analyzed 65 influencers across 81 channels, she argues that such influencers facilitate radicalization through social networking: by referencing others in the network, and through guest appearances on each other's

shows, they let audiences move from mainstream conservative to more and more extreme right-wing contents. She explains their appeal by describing how they establish an alternative sense of credibility based on personal authenticity and relatability, and cultivating a social identity of a countercultural underdog.

Over the past two years, social media platforms have started adjusting their recommender algorithms in a bid to suppress harmful content. A key question going forward is whether these changes will have their intended effect, and what other (unintended) consequences they might have for online discourse.

Analysis: How are platforms and governments addressing the algorithmic amplification of hate speech and disinformation?

Recommendation systems are currently subject to a range of efforts to combat the amplification of disinformation and hate speech. These include (proposed) government interventions, as well as self-regulatory initiatives among platforms and/or civil society. Generally speaking, their toolbox includes the following interventions: content removal and other forms of moderation, algorithmic content curation, user customization options, transparency, and media literacy.

Content removal, demonetization, and/or downranking

One way to respond to amplification concerns is to remove the content at issue. Of course, such content moderation is by no means straightforward. Content removal is not always effective at combating the perceived harms or identifying the content at issue, and raises questions about gatekeeping and the freedom of expression. When it comes to amplification, content removal is especially limited because it may concern content that does not necessarily violate content policies. For instance, one might be concerned about the lack of political balance on a given recommendation system without wishing to actually prohibit content from a particular political orientation or origin. Removal, in other words, is a relatively extreme measure and may not always be proportional.

Platform moderation consists of removing dubious actors and content. In theory, this can help to stop the spread of disinformation, hate speech, and violent extremism. In practice, this is a complex exercise: how are these takedown decisions implemented, following what rules and procedures? Can the speakers object or intervene, and if so how? How are human rights (freedom of speech, rights to information) guaranteed when private actors (platforms) have such censoring powers? How are critical voices not marginalized? Are platforms in any way accountable for their moderation practices, and how can or should their decision-making be publicly scrutinized? Research shows that content takedown is not always effective, either: much of the harmful content and actors remain active, giving them a false aura of legitimacy.[83]

Content moderators have other tools than removal. For instance, YouTube demonetizes some content, terminating or suspending any revenue sharing agreements with the content provider. This can be a powerful deterrent, since many YouTubers rely on ad revenue as a key source of income, and raises comparable concerns from a free speech perspective, at least for professional content producers. By cutting off a key source of revenue, platforms can render certain content unviable to produce, and effectively silence certain speakers.

Another alternative to content removal is downranking: the harmful content is deprioritized in news feeds and other recommendation systems, so that it becomes less visible and less likely to be amplified. This is where the content moderation debate intersects with the broader issue of algorithmic content curation: what design principles are applied in recommendation systems, and how does it pick winners and losers? This is discussed in detail below.

Algorithmic content curation (and non-discrimination)

As discussed, recommendation systems are not inherently neutral and are designed to prioritize content with certain characteristics and deprioritize other content. As such, their functioning is already a form of content moderation: by suggesting some types of content and hiding others, they perform an important gatekeeping function. A number of governments are now proposing to have platform recommendation algorithms accommodate public interest considerations and legal requirements, and platforms are taking comparable measures on their own initiative. For instance, the European Commission's Code of Practice on Disinformation requires platforms to "[d]ilute the visibility of disinformation by improving the findability of trustworthy content" and to "invest in technological means to prioritize relevant, authentic, and authoritative information."[84] The Council of Europe emphasizes the importance of diversity. Its Committee of Ministers has called on member states to foster partnerships between social media platforms and outside stakeholders "to enhance users' effective exposure to the broadest possible diversity of media content."[85] Many of these "public interest considerations," however, are relatively undertheorized and underdeveloped, in terms of providing guidance that can be operationalized and evaluated in algorithmic systems.

Platforms are taking comparable measures of their own accord. In 2018 alone, Facebook announced updates to promote content from friends and reduce the reach of news pages; to downrank "false news" content flagged by accredited fact-checkers; and to downrank "borderline content" that falls short of violating company policies.[86] In April 2019, Facebook started downranking anti-vaccination content.[87] In May 2019, Facebook presented its new "click-gap" method to suppress "low-quality content."[88] Similarly, in January 2019, YouTube announced that it would "begin reducing recommendations of borderline content and content that could misinform users in harmful ways."[89]

Another set of efforts focuses instead on non-discrimination rules, which would place limits on such algorithmic curation. For instance, the German federal broadcasting authority has proposed, in an instrument known as the *Medienstaatsvertrag,* to prohibit platforms from discriminating against "journalistic editorial content" to the extent that the intermediary has "potentially a significant influence on their visibility."[90] In the Netherlands, the Dutch State Commission on the Parliamentary System proposed a comparable "independent entity" to monitor platform recommendations, but unlike the Germans, its mandate would not focus on non-discrimination but rather on maintaining "diversity" and avoiding "bias."[91] In the U.S., Senator Josh Hawley (R-MO) has proposed that platforms observe "political balance" in their algorithms.[92] However, implementing such principles in practice is not straightforward: as discussed, it is not clear how recommendation algorithms can be made "neutral," or what would constitute "discrimination," since they necessarily rank some content over others. In some sense, the entire purpose of these algorithms is to discriminate. Even more difficult are concepts like "political balance," "bias," and "diversity," which implicate value-laden and content-specific judgements about newsworthiness. Indeed, these two types of principles

(non-discrimination and diversity) may contradict each other; diversity rules could require platforms to prioritize certain specific types of content, whereas non-discrimination rules might prohibit them from doing so.

Without further elaboration, therefore, these regulatory standards seem immature at best, and implementing them would be both technically infeasible and politically controversial. Particularly with such vague and subjective rules, government action in this space also raises questions about freedom of expression and the rule of law. Without adequate safeguards, government regulation of content recommendation could impede the freedom of expression of social media users. Prescribing what should be downranked risks becoming a form of censorship, and what must be prioritized a form of propaganda.

Finally, it is worth noting that antitrust and consumer protection law already place some limits on algorithmic curation and discrimination. First, antitrust law could place limits on the ability of platforms to prioritize their own services (as seen in the EU case against Google's Search and Shopping products). Secondly, consumer protection law in most jurisdictions forbids covert advertising, which means that platforms have a duty to disclose whether content is being sponsored – i.e., whether it constitutes an advertisement rather than organic content. Such disclosures are already common practice on most major platforms, but they are not as visible and recognizable as they could be: advertisements are often incorporated into organic content feeds, and their design risks blurring the boundaries between these two types of content.

As noted, platforms also respond to public pressure and changed their recommendation algorithms to prioritize "trusted sources" (whitelisting) and deprioritize "harmful content" (blacklisting), which runs into the same problems as content moderation.[93] Downranking also raises many of the same free speech issues as content moderation: it prevents platform users from effectively making their voices heard, with little to no accountability when their content is removed. Because of "proprietary reasons," platforms are not transparent about what such changes in their recommendation algorithms look like, hindering public scrutiny and accountability. Some argue that as long as platforms work on ad-based models, they have an incentive to permit, or at least minimize, their investments in combating disinformation and other incendiary content because this content keeps people engaged on their platforms and thereby creates a source of revenue.[94] Public pressure may still motivate platforms to act (or be seen to act), but these countervailing incentives should be taken into account when assessing the prospects of self-regulation in this space.

User customization options

Another policy option is to develop tools for user choice, so they can customize their recommendation systems. The aforementioned Council of Europe recommendation calls on states to encourage platforms to "provide clear information to users on how to find, access and derive maximum benefit from the wide range of content that is available."[95] Similar rules are found in the European Commission's Code of Practice and the German broadcaster's proposals.

In practice, platforms already provide several options for user customization. Basic functionalities such as liking, following, and subscribing can be seen as a form of user choice, since they help users to express what kind of content they would like to receive. YouTube and Facebook allow users to

express *disinterest* and filter out certain kinds of content or sources. Twitter allows users to view tweets in a chronological order to avoid additional algorithmic curation. However, users may not always be aware of or interested in these features – particularly when they are not easily visible and accessible in the platform's visual interface.

Transparency

Perhaps the most widely supported policy priority around recommendation systems is transparency. All of the aforementioned policy instruments include transparency obligations of some sort. Further transparency obligations can also be found in horizontal instruments. The General Data Protection Regulation grants data subjects the right to demand "explanations" about recommendation systems.[96] Similarly, the EU's Regulation on Promoting Fairness and Transparency for Business Users of Online Intermediation Services (Platform-to-Business Regulation) requires platforms to explain "the characteristics of the goods and services offered to consumers through the online intermediation services or the online search engine." The revised Audiovisual Media Services Directive has additional rules specifically for video platforms. However, transparency is a broad term and can take many forms -- particularly in the context of technically complex systems like content recommendations. These issues of transparency are given particular attention in a separate Transatlantic Working Group document.[97]

Media literacy

Last but not least, interventions can target users themselves: improving media literacy. Most, if not all, European states have programs, often with government support, aimed at increasing people's ability to discern information quality. Such programs can include practical skills, critical reasoning, and ethical considerations. The idea is simple: education better enables people to navigate today's complex media ecosystems and to be more resilient against the propaganda campaigns of malicious actors. Media literacy is recognized as a fundamental skill in the 2016 Audiovisual Media Services Directive, a key document defining the EU's future media and communications policy. The European Commission supports initiatives and prizes, manages projects, programs, and funding schemes, such as Creative Europe, coordinates with member states on policies and best practices, and develops new policies based on expert group findings. Each individual member state has similar programs that tackle this issue from country-specific perspectives, often in conjunction with civil society and educational initiatives.

Raising awareness and media literacy is useful and important; because these automated systems are both novel and complex, they are poorly understood by many users. Helping them to understand how recommendations are generated and what they can do to alter and customize their experience can help users to act with greater care and autonomy in the face of harmful or misleading content offerings. Still, this approach is limited by the fact that some people willfully engage in spreading contentious content. Sociological research shows that these people are attracted by disinformation not necessarily because they consider it truthful, but rather because it aligns with their worldview and it gives them a sense of community and identity.[98] A related problem is that those in most need of "education" may not be the ones who are actually reached and most receptive to it. Therefore, educating citizens may not always work in combating disinformation in society at large because the issue is not just cognitive – i.e., based on a faulty understanding of social media or recommender

algorithms – but is also driven by deeper and more complex cultural and societal shifts related to the loss of trust in mainstream media, science, and other knowledge institutions.

Fact-checking

Tracing, highlighting, and correcting disinformation is important for reasons of transparency and truth-finding, but is not a comprehensive solution either. At a practical level, fact-checking is difficult to perform at scale: determining truth is one of the most difficult aspects of content moderation, even for trained fact-checkers.[99] Even if possible, it will be properly difficult to automate and it will continue to require human review, which is costly and time-consuming.[100] More fundamentally, as discussed, the truthfulness of disinformation is often of secondary importance to those producing and sharing it: it is mainly about worldview and identity, not truth. Accordingly, fact-checking corrections will not always be taken seriously by people who ideologically do not align with its findings.[101] Moreover, because fact-checkers are often from opposed ideological and societal groups, these organizations may suffer a lack of trust among their target audience.[102] Indeed, highlighting false information may even be counterproductive, since the excessive attention (and uptake by other media organizations) can also increase its reach.[103] It has even been argued that being confronted with corrections could actually further *strengthen* the original beliefs.[104]

(Self) Regulation

Various regulatory frameworks have been discussed to tackle the amplification of disinformation and hate speech, ranging from aforementioned efforts to impose more diverse recommendations to stricter rules for content moderation.[105] However, governments have been reluctant to take a strong lead here. For now, (supra)national governments have tried some forms of co-regulation, such as the 2018 European Commission's Code of Practice, in which the major tech companies pledged to work more actively to lessen the spread of disinformation and hate speech online. However, all of this is non-binding and the rules of the game can be rather freely interpreted by these companies. Various critics in academia and civil society argue that such forms of self-regulation do not provide enough incentives for platforms to make meaningful changes, which might threaten their core business models. Fact-checking organizations and academic researchers who partnered with Facebook have expressed dissatisfaction at the lack of progress in these arrangements. Some argue therefore to regulate recommendation systems themselves.[106] The development of effective regulations may be challenging, not only due to its technical complexity and dynamism but also due to its extreme political sensitivity, and its implications for fundamental rights including freedom of expression.

Recommendations

Despite these difficulties in preventing the amplification of disinformation and hate speech online, it remains important to think about other possible remedies and solutions. We propose here some potential directions and evaluate their feasibility.

Raising awareness and transparency

While transparency of recommendation systems is surely no panacea, and it knows many pitfalls,[107] there is much to gain by raising awareness of their functioning. Transparency can help to hold these systems accountable and enable more evidence-based policy making. To this end, governments could

impose stricter enforcement of users' right to explanation under the GDPR, and require platforms to offer additional forms of transparency of their recommendation systems. This could take several forms, including user-facing notices, government or civil society auditing, academic partnerships, and regimes for public disclosure (discussed in the TWG's working paper on transparency).

One particularly important form of transparency is the sharing of data and information with outside researchers. First, non-sensitive, anonymized data could be shared in public datasets.[108] Second, sensitive data can be shared in partnerships with relevant institutions, under non-disclosure disagreements to safeguard confidentiality. Given the technical and legal difficulties faced by self-regulatory initiatives in this space, such as Social Science One, there may be a role for governments to facilitate these exchanges (e.g., by providing a processing ground under data protection law, and/or by imposing sanctions for breaches of confidentiality).

To improve awareness, legacy media and civil society organizations should pay more attention to the social and cultural contexts in which people radicalize, rather than just criticizing social media platforms. Reporting on recommendation systems should not lose sight of the bigger picture. For instance, the Mozilla Foundation took up an article by the Washingtonian, which reported on a mother and her teenage son who turned to white supremacist movements online after being falsely accused of a sex offense at high school.[109] Mozilla brought the mother and son together and invited their community to pose questions to them. Personal stories like these help our understanding about radicalization processes beyond a sole focus on the radicalizing powers of recommendation systems.

Increasing user control

Some scholars argue that increasing user control in recommendation systems would mitigate many of the aforementioned problems, and enhance the individual and societal value of recommendation systems.[110] Besides empowering users to make recommendation systems more responsive to *their* interests and needs, and not what platforms think is most relevant to them, user control simultaneously requires recommendation systems to be more transparent and explainable. Not only can this improve user satisfaction and trust, it can also counteract "filter bubble" concerns by encouraging them to look beyond their assumed or known interests. Users could, for example, be confronted with explicit options to receive recommendations outside their ordinary consumption habits, like a "get me out of my filter bubble button" or a "show me more of the ideological other side" slider. There are several ways to increase user control, at the input level of user preference and in choosing from different kinds of recommendation algorithms so as to get different *kinds* of recommendations depending on one's mood or information interests at a given moment.[111] This being said, some would warn that increasing user control can also serve to enable users who deliberately choose to view extremist or contentious content. Much depends on how user control is implemented and designed, and more empirical research (and access to data) is needed to study the effects of these tools in practice.

Multistakeholder governance

In designing and implementing these interventions, one way to mitigate free speech concerns regarding social media regulation is to incorporate multistakeholder and co-regulatory elements.[112] For instance, the Council of Europe's guidelines related to media pluralism policy recommend that

social media platforms and other online services engage in "open, independent, transparent, and participatory initiatives" alongside "media actors, regulatory authorities, civil society, academia and other relevant stakeholders" on such issues as algorithmic diversity and transparency.[113] Such initiatives could assist in many of the efforts discussed above, including the design of recommender systems and algorithmic curation as well as interventions for transparency and media literacy.

Bringing different stakeholders together, instead of relying exclusively on the judgment of either dominant platforms or government regulators, can create checks and balances on these powerful actors while helping them take account of outside viewpoints, interests, and expertise. In this way, opening up the governance process to experts and affected stakeholders has the potential to enhance both its effectiveness and its legitimacy.[114]

Designing effective and inclusive regulatory institutions of participation is a key challenge going forward; on the one hand, strictly voluntary arrangements may fail to hold powerful commercial actors accountable. Government, therefore, may have an important role to play in undergirding these initiatives with the force of law. On the other hand, the participation of outside experts should not serve as a mere fig leaf on government policy; to have its intended effect, multistakeholder governance should meaningfully involve other actors in the decision-making process. These challenges may be serious, but inspiration can be drawn from earlier precedents in media governance; for instance, lessons can be drawn from the design of self-regulatory and co-regulatory bodies in journalism, advertising, and public broadcasting. Another relevant avenue of debate is the concept of "social media councils," explored in a separate Working Group report.[115]

Incentives for alternative business models

As discussed, ad-based business models give social media platforms a strong commercial incentive to keep users engaged, including via controversial and incendiary contents. Some critics have therefore argued to push or incentivize platforms toward a subscription-based model.[116] Subscription-based models have the advantage of a revenue stream irrespective of advertising, meaning that recommendation systems need not be designed to optimize for engagement only. Netflix and Spotify, the two largest subscription-based platforms, increasingly use other metrics, such as user satisfaction and explorability as well. Although such an approach introduces social equity issues (people with lower incomes may not be able or willing to pay for ad-free services), it could thus make good sense to think of financial or regulatory incentives to push ad-based platforms to other business models with fewer incentives to push inflammatory content. One starting point, for instance, might be to impose organizational restrictions to reduce the influence of advertising operations' on organic content recommendations. Lessons could be drawn from best practices in journalism, where editorial decision-making is also designed to be independent from commercial operations. These are evidently more far-reaching interventions than the aforementioned, and have not yet received as detailed a treatment in policy research and debate. Further research would therefore be needed to explore the potential applications and viability of these more muscular approaches to social media regulation.

Conclusion

Platforms use countless different forms of automation to shape our experiences online. As we have seen, this is not just about "AI." And although technological solutions for content moderation and curation are increasingly widespread, many are still rudimentary and imperfect. Use of automation in content moderation exposes all speech to a form of evaluation *ex ante* and in a way that fails to consider linguistic, social, historical, and other relevant context – which creates substantial risks to the freedom of expression. Governments should therefore act with caution and resist simplistic narratives about all-powerful algorithms or AI as being the sole cause of, or solution to, the spread of harmful content online. Indeed, any legal requirements to adopt specific forms of automation are likely to be premature, and would present major risks to freedom of expression. For now, what governments should focus on is enhancing transparency in existing practices, empowering research communities with the necessary data, and ensuring that users have access to meaningful choice and redress mechanisms.

Notes

[1] Emma Llansó is director of the Center for Democracy & Technology's Free Expression Project, which works to promote law and policy that support internet users' free expression rights in the United States and around the world.

[2] Joris van Hoboken is a senior researcher at the Institute for Information Law (IViR), and a professor of law at the Vrije Universiteit Brussels (VUB). He works on the intersection of fundamental rights protection (data privacy, freedom of expression, non-discrimination) and the governance of platforms and internet-based services.

[3] Paddy Leerssen is a Ph.D. candidate at the Institute for Information Law (IViR). He focuses on the European governance of recommendation algorithms in social media platforms and their impact on media pluralism – how platforms can be made publicly accountable

[4] Jaron Harambam is an interdisciplinary sociologist working on news, disinformation, and conspiracy theories in today's algorithmically structured media ecosystem. He currently holds a Marie Sklodowska-Curie Individual Fellowship at KU Leuven's Institute for Media Studies.

[5] For example, one set of estimates from 2018 found that, every minute, users post nearly 2,000 comments on Reddit, 50,000 photos on Instagram, 80,000 posts on Tumblr, 470,000 tweets on Twitter, 2 million snaps on Snapchat, and 3.8 million search queries on Google, Domo, Data Never Sleeps 6.0, https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf.

[6] A note on terminology: The phrase "artificial intelligence" broadly refers to computer systems that can perform tasks associated with intelligent beings. Its use in policy-making processes is typically imprecise, evoking powerful technical capabilities without necessarily specifying existing technical processes. "Machine learning" is a branch of computer science focused on computer programs that adapt to and "learn" how to act from data without being specifically programmed by a human.

[7] Use of AI in Content Moderation, Cambridge Consultants for UK OfCom (2019) at p. 37, https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

[8] *See* Ekram Sabir et al., Recurrent Convolutional Strategies for Face Manipulation Detection in Videos, https://arxiv.org/abs/1905.00582.

[9] This use of unsupervised learning to create labeled inputs for a supervised-learning process is sometimes called "self-supervised" learning. *See* Andriy Burkov, https://www.kdnuggets.com/2019/01/burkov-self-supervised-learning-word-embeddings.html.

[10] E.g. T. Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (July 2016), https://arxiv.org/abs/1607.06520.

[11] https://towardsdatascience.com/the-a-z-of-ai-and-machine-learning-comprehensive-glossary-fb6f0dd8230

[12] https://developers.google.com/machine-learning/glossary#recurrent_neural_network

[13] Mixed Messages p. 9, https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

[14] "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," Felbo et al., https://arxiv.org/pdf/1708.00524.pdf

[15] https://www.perspectiveapi.com/#/home

[16] https://github.com/conversationai/perspectiveapi/wiki/perspective-hacks

[17] https://docs.coralproject.net/talk/toxic-comments/

[18] Deceiving Google's Perspective API Built for Detecting Toxic Comments, Hosseini et al. (2017) https://arxiv.org/pdf/1702.08138.pdf.

[19] The Risk of Racial Bias in Hate Speech Detection, Sap et al. (2019), https://www.scribd.com/document/421898931/The-Risk-of-Racial-Bias-in-Hate-Speech-Detection

[20] https://conversationai.github.io/

[21] https://openai.com/blog/better-language-models/. Specifically, "In order to preserve document quality, we used only pages which have been curated/filtered by humans—specifically, we used outbound links from Reddit which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting (whether educational or funny), leading to higher data quality than other similar datasets, such as CommonCrawl."

[22] Language Models are Unsupervised Multitask Learners, Radford et al. (2019) https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. See also https://nostalgebraist.tumblr.com/post/187579086034/it-seems-pretty-clear-to-me-by-now-that-gpt-2-is

[23] https://openai.com/blog/better-language-models/

[24] Release Strategies and the Social Impacts of Language Models, Solaiman et al. (2019) https://arxiv.org/ftp/arxiv/papers/1908/1908.09203.pdf. *See also*, OpenAI has released the largest version yet of its fake-news-spewing AI, Karen Hao, (Aug. 29, 2019), https://www.technologyreview.com/s/614237/openai-released-its-fake-news-ai-gpt-2/

[25] OpenGPT-2: We Replicated GPT-2 Because You Can Too, Aaron Gokaslan and Vanya Cohen et al., (Aug. 26, 2019) https://blog.usejournal.com/opengpt-2-we-replicated-gpt-2-because-you-can-too-45e34e6d36dc

[26] If the goal is to ensure the integrity of the message, as in cryptographic hashing, then this sensitivity of the hash to the most minute change is the goal of running the hash function. Perceptual image hashes are much more vulnerable to attacks that can reveal information about the content that has been hashed. *See* https://towardsdatascience.com/black-box-attacks-on-perceptual-image-hashes-with-gans-cc1be11f277

[27] *See* Perceptual video hashing based on the Achlioptas's random projections, R. Sandeep and Prabin K. Bora, https://ieeexplore.ieee.org/document/6776252

[28] https://jenssegers.com/perceptual-image-hashes

[29] https://web.archive.org/web/20130921055218/http://www.microsoft.com/global/en-us/news/publishingimages/ImageGallery/Images/Infographics/PhotoDNA/flowchart_photodna_Web.jpg

[30] https://medium.com/@timanglade/how-hbos-silicon-valley-built-not-hotdog-with-mobile-tensorflow-keras-react-native-ef03260747f3

[31] https://gizmodo.com/british-cops-want-to-use-ai-to-spot-porn-but-it-keeps-m-1821384511

[32] Note the difference between face detection, which identifies the presence of (typically human) faces in an image, and facial recognition, which attempts to match a detected face to an identified person.

[33] Louise Matsakis, Tumblrs Porn-Detecting AI Has One Job--And It's Bad At It, https://www.wired.com/story/tumblr-porn-ai-adult-content/ (Dec. 5, 2018).

[34] OfCom 2019, 51.

[35] For example, Facebook has announced plans to develop a dataset for researchers to use in developing technology to detect deepfakes. Facebook, "Creating a dataset and a challenge for deepfakes" (Sept. 5, 2019) https://ai.facebook.com/blog/deepfake-detection-challenge/

[36] *See* Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, Explaining and Harnessing Adversarial Examples (2015) https://arxiv.org/pdf/1412.6572v3.pdf (example of image transformation that leads classifier to conclude image of panda is a gibbon).

[37] Nicholas Papernot et al., Practical Black-Box Attacks against Deep Learning Systems Using Adversarial Examples (2016), https://arxiv.org/pdf/1602.02697v2.pdf

[38] https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed

[39] See Adrian Yijie Xu, AI, Truth, and Society: Deepfakes at the front of the Technological Cold War, (July 2, 2019) https://medium.com/gradientcrescent/ai-truth-and-society-deepfakes-at-the-front-of-the-technological-cold-war-86c3b5103ce6

[40] Id.

[41] https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography

[42] Joe Uchill, Why the deepfakes threat is shallow (Aug. 15, 2019), https://www.axios.com/why-the-deepfakes-threat-is-shallow-16caf6a0-af83-4dbc-9008-6a2d4a2f08ae.html

[43] *See,* Camille François, Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses, https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf

[44] *See* Mixed Messages, *supra* n.13, p.14-15.

[45] http://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/

[46] OfCom 2019, p. 26.

[47] See, e.g., Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx

[48] See, e.g., the recent Declaration (Feb. 13, 2019)1 on the manipulative capabilities of algorithmic processes. The Council of Europe is finalizing a round of consultations on its draft recommendation on the human rights impacts

[49] Ananny, M. (2019). Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance. *Free Speech Futures, An essay series reimagining the First Amendment in the digital age*, Columbia University. ("Today, the meaning and force of the First Amendment play out in the new and often unstable technological infrastructures and institutional spaces of social media platforms.")

[50] Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx

[51] *See* Joris van Hoboken and Daphne Keller, Design Principles for Intermediary Liability Laws, TWG Discussion paper, October 2019, available at https://www.ivir.nl/publicaties/download/Intermediary_liability_Oct_2019.pdf.

[52] *See*, for instance, Article 17 of the EU's new 'Copyright Directive': Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. OJ L 130, available at: http://data.europa.eu/eli/dir/2019/790/oj.

[53] *See, for example*, Mark McCarthy, Transatlantic Working Group paper

[54] *For example*, Google conducted, and published, a human rights impact assessment for its Celebrity Recognition API (October 2019): https://services.google.com/fh/files/blogs/bsr-google-cr-api-hria-executive-summary.pdf. Facebook has since published a human rights impact assessment of its newly formed Oversight Board: https://bsr.app.box.com/s/8r0vw4a5kib6y6xfdtd5jin3g3fcbzs5.

[55] *See* Part II Explaining recommendation systems and their deployment below (Under "The role of recommendation systems in online hate speech and disinformation").

[56] Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *Arxiv 1904.02095v4 [Cs]*. Retrieved from https://arxiv.org/pdf/1904.02095.pdf

[57] Ricci, F., Rokach L., & Shapira, B. (2015) *Recommender Systems Handbook*. Springer, Cham, SH.

[58] Ibid.

[59] 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.

[60] Diakopoulos, N. (2015) Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3): 398-415.

[61] Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133).

[62] Burrell, J. (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3: 1.; Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

[63] For instance, Google provides relatively detailed guidance to 3rd party reviewers that evaluate search results:: https://www.google.com/search/howsearchworks/mission/users/ and https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

[64] *See* Part II Analysis: How are platforms and governments addressing the algorithmic amplification of hate speech and disinformation? below (Under 'Algorithmic content curation (and non-discrimination').

[65] Cobbe, J., & Singh, J. (2019). Regulating Recommending: Motivations, Considerations, and Principles. (April 15, 2019). Available at SSRN: https://ssrn.com/abstract=3371830 or http://dx.doi.org/10.2139/ssrn.3371830; Gary, J., and Soltani, A. (2019) *First Things First: Online Advertising Practices and Their Effects on Platform Speech*, Knight First Amendment Institute at Columbia University, available at: https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech; Lewis, P. (2018). Fiction is outperforming reality": How YouTube's algorithm distorts truth. *The Guardian*, February 2, 2018. A recent study on YouTube's algorithms refutes the radicalization claim and finds evidence that YouTube's recommendation algorithm favors mainstream sources. See Ledwich, M., & Zaitsev, A. (2019). Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization. *arXiv preprint arXiv:1912.11211*.

[66] *See* M Golebiewski and D Boyd, Data Voids: Where Missing Data Can Easily Be Exploited (2018), https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

[67] Bahara, H., Kranenberg, A., Tokmetzis, D. (2019) Hoe YouTube rechtse radicalisering in de hand werkt, *De Volkskrant*, 8 februari 2019.

[68] *See* M. Golebiewski and D. Boyd, Data Voids: Where Missing Data Can Easily Be Exploited (2018), https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

[69] Napoli, P. (2014). Digital intermediaries and the public interest standard in algorithm governance. *Media Policy Blog*.

[70] Bennet and Livingston, 2018

[71] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353-369.

[72] *See* Helberger, N. (2019, in press). On the democratic role of news recommenders, *Digital Journalism*. Available at: https://www.tandfonline.com/doi/full/10.1080/21670811.2019.1623700

[73] Cadwalladr, C., & Graham-Harrison, E. (2018). The Cambridge Analytica files. *The Guardian*, 21, 6-7.

[74] Moore, M., & Tambini, D. (2018). *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*. Oxford University Press.

[75] Helberger et al., 2018

[76] Lewis, P. & McCormick, E. (2018). How an ex-YouTube insider investigated its secret algorithm. *The Guardian*, February 2, 2018.

[77] Lewis, 2018

[78] Tufekci, 2018

[79] Lewis, 2018

[80] Kaiser, J. & Rauchfleisch, A. (2019) The implications of venturing down the rabbit hole, *Internet Policy Review*, June 27 2019

[81] Serrato, R. (2018) How YouTube's algorithm amplified the right during Chemnitz, Algorithmic Accountability Reporting at AlgorithmWatch, Berlin, November 5 2018.

[82] Bahara et al., 2019

[83] Gary & Soltani, 2019

[84] European Commission (2018). *Code of Practice on Disinformation.* Available at: https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

[85] Council of Europe, 2018. Recommendation CM/Rec(2018)1[1] of the Committee of Ministers to member States on media pluralism and transparency of media ownership. Available at https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680790e13.

[86] Mosseri, A. (2018). Bringing People Close Together. *Facebook Newsroom.* Available at: https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/; Facebook (2018). How People Help Fight False News. *Facebook Newsroom.* Available at: https://newsroom.fb.com/news/2018/06/inside-feed-how-people-help-fight-false-news/; Zuckerberg, M. (2018). A Blueprint for Content Governance and Enforcement. *Facebook Notes*. Available at: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

[87] Bickert, M. (2019). Combatting Vaccine Misinformation. *Facebook Newsroom*. Available at: https://newsroom.fb.com/news/2019/03/combatting-vaccine-misinformation/.

[88] Rosen, G. (2019). Remove, Reduce, Inform: New Steps to Manage Problematic Content. *Facebook Newsroom*. Available at: https://newsroom.fb.com/news/2019/04/remove-reduce-inform-new-steps/.

[89] YouTube, 2019

[90] Helberger, Leerssen & Van Drunen (2019). Germany proposes Europe's first diversity rules for social media platforms. *LSE Media Policy Project Blog.* Available at: https://blogs.lse.ac.uk/mediapolicyproject/2019/05/29/germany-proposes-europes-first-diversity-rules-for-social-media-platforms/.

[91]https://www.staatscommissieparlementairstelsel.nl/documenten/rapporten/samenvattingen/12/13/eindrapport

[92] https://www.vox.com/2019/6/26/18691528/section-230-josh-hawley-conservatism-twitter-facebook

[93] Cobbe & Singh, 2019; Gary & Soltani, 2019

[94] Ibid.

[95] Council of Europe, 2018, par. 2.5.

[96] Van Drunen, Helberger & Bastian (2019). Know your algorithm: what media organizations need to explain to their users about news personalization. *International Data Privacy Law* 9(3). https://academic.oup.com/idpl/advance-article/doi/10.1093/idpl/ipz011/5544759.

[97] MacCarthy (2020). Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry.

[98] Harambam, J. (2017) "The Truth Is Out There": Conspiracy culture in an age of epistemic instability. Rotterdam: Erasmus University.

[99] Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.; Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 1-22.; Swire et al., 2017; Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

[100] Graves, L. (2018). Understanding the promise and limits of automated fact-checking. *Factsheet*, *2*, 2018-02.

[101] Graves, L. (2016).

[102] Harambam, J. (2017b). De/politisering van de Waarheid. *Sociologie*, *13*(1), 73-92.

[103] Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*(6), 701-723.; Lewandowksy et al., 2017

[104] Strandberg, K., Himmelroos, S., & Grönlund, K. (2019). Do discussions in like-minded groups necessarily lead to more extreme opinions? Deliberative democracy and group polarization. *International Political Science Review*, *40*(1), 41-57.; Wojcieszak, M. (2011). Deliberation and attitude polarization. *Journal of Communication*, *61*(4), 596-617.

[105] Cobbe & Singh, 2019

[106] Id.

[107] 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 Law & Technology Review*, *16*, 18.

[108] The technical challenge, however, of anonymizing datasets should not be underestimated. *See* Elizabeth Gibney, Privacy hurdles thwart Facebook democracy research https://www.nature.com/articles/d41586-019-02966-x?sf220739510=1. *See also* Kobbi Nissim et al., Differential Privacy: A Primer for a Non-technical Audience, https://privacytools.seas.harvard.edu/files/privacytools/files/nissim_et_al_-_differential_privacy_primer_for_non-technical_audiences_1.pdf

[109] Caltrider, J. Journey From the Dark Side: How One Teen Boy Got Radicalized Online and Came Out the Other Side. *Mozilla Foundation* (2019, August). https://foundation.mozilla.org/en/blog/journey-dark-side/

[110] Harambam, J., Helberger, N., & van Hoboken, J. (2018). Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180088.

[111] Harambam, J., Bountouridis, D., Makhortykh, M., & Van Hoboken, J. (Sept. 2019). Designing for the better by taking users into account: a qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 69-77). ACM.

[112] Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The information society*, *34*(1), 1-14.

[113] Committee of Ministers Recommendation CM/Rec (2018)1 on media pluralism and transparency of media ownership. Council of Europe 2018.

[114] Helberger, et al., 2018

[115] *See* McCarthy, Mark (2020). Transparency Requirements for Digital Social Media Platforms Survey and Recommendations for Policy Makers and Industry.

[116] Gary & Soltani, 2019