Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 4, Pages 222–233
https://doi.org/10.17645/mac.v9i4.4062

Article

Automated Trouble: The Role of Algorithmic Selection in Harms on Social Media Platforms

Florian Saurwein [1,*] and Charlotte Spencer-Smith [2]

[1] Institute for Comparative Media and Communication Studies, Austrian Academy of Sciences, University of Klagenfurt, Austria; E-Mail: florian.saurwein@oeaw.ac.at
[2] Department of Communication Studies, Paris Lodron University of Salzburg, Austria;
E-Mail: charlotte.spencer-smith@sbg.ac.at

* Corresponding author

Submitted: 25 January 2021 | Accepted: 3 June 2021 | Published: in press

Abstract
Social media platforms like Facebook, YouTube, and Twitter have become major objects of criticism for reasons such as privacy violations, anticompetitive practices, and interference in public elections. Some of these problems have been associated with algorithms, but the roles that algorithms play in the emergence of different harms have not yet been systematically explored. This article contributes to closing this research gap with an investigation of the link between algorithms and harms on social media platforms. Evidence of harms involving social media algorithms was collected from media reports and academic papers within a two-year timeframe from 2018 to 2019, covering Facebook, YouTube, Instagram, and Twitter. Harms with similar casual mechanisms were grouped together to inductively develop a typology of algorithmic harm based on the mechanisms involved in their emergence: (1) algorithmic errors, undesirable, or disturbing selections; (2) manipulation by users to achieve algorithmic outputs to harass other users or disrupt public discourse; (3) algorithmic reinforcement of pre-existing harms and inequalities in society; (4) enablement of harmful practices that are opaque and discriminatory; and (5) strengthening of platform power over users, markets, and society. Although the analysis emphasizes the role of algorithms as a cause of online harms, it also demonstrates that harms do not arise from the application of algorithms alone. Instead, harms can be best conceived of as socio-technical assemblages, composed of the use and design of algorithms, platform design, commercial interests, social practices, and context. The article concludes with reflections on possible governance interventions in response to identified socio-technical mechanisms of harm. Notably, while algorithmic errors may be fixed by platforms themselves, growing platform power calls for external oversight.

Keywords
algorithmic content curation; algorithmic harm; algorithms; behavioural advertising; content moderation; internet; social media

Issue
This article is part of the issue “Algorithmic Systems in the Digital Society” edited by Sanne Kruikemeier (University of Amsterdam, The Netherlands), Sophie Boerman (University of Amsterdam, The Netherlands) and Nadine Bol (Tilburg University, The Netherlands).

1. Introduction

In recent years, internet platforms have gained enormously in reach and influence among the broader population. Scholars have therefore pointed to a trend towards platformisation and the development of a platform society (Helmond, 2015; van Dijck et al., 2018). Among platform services, social media enjoy particular popularity: 2,7 billion people now use at least one of the applications owned by Facebook (Facebook, Instagram, WhatsApp, Messenger), of which Facebook alone has 1,84 billion daily active users (Facebook

Investor Relations, 2021). However, the rise of social media platforms has also attracted strong criticism due to a range of social and economic harms. This includes privacy violations through collection and processing of user data, the potential for social sorting and discrimination, the growth of surveillance capitalism (Zuboff, 2019), and the promotion of intense and addictive user behaviour. Critics point to the consequences of growing platform power (e.g., Helberger, 2020), as internet platforms exert significant influence over societal communication and can strategically use their technologies to direct user attention and shape reality. Platforms stand accused of harming public discourse and democracy by fuelling social fragmentation, political bias, and polarisation, and by contributing to the spread of problematic content such as hate speech and disinformation (Persily & Tucker, 2020). Moreover, platforms are criticised for perpetuating economic harms, such as increasing the dependence of sectors like retail and publishing on platform intermediaries, evading tax, and abusing dominant market positions, which has already led to antitrust cases and fines in Europe and the US (US House Judiciary Subcommittee on Antitrust, Commercial, and Administrative Law, 2020). In addition to this, a long-standing debate continues over whether platforms damage economic rights by enabling copyright violations or damage freedom of expression by enforcing copyright too heavy-handedly.

In particular, algorithms have been identified as a contributing factor in a number of these harms. Indeed, several applications on social media platforms are based on algorithms. Algorithms are used to perform functions such as monitoring, scoring, recommendation, forecasting, and automated transactions. Platforms use algorithms to provide personalised news feeds, features on "trending topics," search and autocomplete functions, computational advertising, contact and group recommendations, as well as to identify and filter unwanted content (e.g., pornography, spam, disinformation). In addition, third parties use their own algorithmic tools on social media platforms, such as chatbots and clickbots. Third parties also deploy algorithms for the purposes of social scoring to offer credit or insurance based on the analysis of the customer's social media posts, and the People's Republic of China is developing a social credit system. In sum, algorithms play a wide range of roles on social media platforms and are believed to contribute to several harms that have emerged with the development of the internet.

When harms arise, it may be tempting to "blame it" on the algorithm, but the link itself between algorithms and harm is often unclear. In addressing problems caused by algorithms, the literature describes harms that range from damage to democratic processes (Tufekci, 2015) to economic harm, which may be genuinely unintentional or purposefully abusive (Muller, 2020). In the context of this article, we use the term "algorithmic harm" to describe harmful or negative effects upon individuals, markets, and society caused in part or in full by the use of algorithms. Based on this definition, the aim of the article is to contribute to the understanding of algorithmic harm by exploring the roles that algorithms play in the emergence of harms on social media platforms. To analyse these roles, the article provides an introductory primer on the use of algorithms on social media platforms, describing their application and purpose in recommendation, content moderation, and advertising. After a brief description of the method, the article goes on to present the results. From a broad collection of case studies of harms, the article develops a typology that distinguishes five areas of harm according to the role that algorithms play in their emergence: Algorithms are (1) deficient tools that lead to errors, (2) instruments that serve manipulation, (3) amplifiers of problematic content, (4) enabling structures for problematic behaviour, and (5) instruments of platform power. This typology contributes to a nuanced understanding of the role of algorithms in the emergence of harms and offers a basis to draw preliminary conclusions for governance strategies to combat algorithmic harms.

2. Areas of Application of Algorithms on Social Media Platforms

In recent years, attention has been increasingly paid to algorithms as they enter more and more areas of public life. On the Internet, algorithms are abstract procedures implemented in software programmes that transform input through specified computational procedures (throughput) into output. Many of these programmes are developed to handle the massive data and information available online (input). They therefore screen and assign relevance to data, select information, and put it into order (throughput). The output may take on different forms to be used in functions such as rankings, recommendations, price setting, or text. Latzer et al. (2016, p. 397) suggest the term "algorithmic selection" to describe these operations, defined as "a process that assigns relevance to information elements of a data set by an automated, statistical assessment of decentrally-generated data signals." The centrepiece of this process model is the throughput stage at which the algorithms operate that define the input–output relationship. Although input, throughput, and output vary for different services, algorithmic selection builds the techno-functional core of a number of internet applications in a broad range of fields and for diverse functions, such as search (e.g., search engines), aggregation (e.g., news aggregators), observation and surveillance (e.g., government surveillance), forecasting (e.g., predictive policing), recommendation (e.g., music platforms), scoring (e.g., credit scoring), content production (e.g., robot journalism), and allocation (e.g., computational advertising). Social media platforms too rely heavily on algorithmic selection. For the purposes of orientation, this article highlights three key areas of application

central to the day-to-day operations of major social media platforms.

2.1. Curation, Recommendation, and Discovery

As users upload content to social media platforms at an incredible rate (Hale, 2019), algorithms help sort through the flood of information to show the most relevant content to each user. Such algorithms take into account the interests, preferences, past behaviour, and predicted behaviour of a particular user to recommend content that might interest them (Cobbe & Singh, 2019). They may also recommend content that is popular among other, similar users, or among all users on the platform. Well-known examples are the Facebook News Feed algorithm (DeVito, 2017) and the YouTube algorithm that selects the next video to play automatically. Such personalised recommendation algorithms help users by showing them relevant content, but they are also engineered in the interests of the platform to maximise user time and engagement on the site (Bergen, 2019). Beyond the feed, algorithms also suggest other users to connect with, pages to follow, or groups to join. Search engines autocomplete functions operate while a user is typing a term into the search engine of a platform, making suggestions by automatically completing the search term. This may reveal terms that other users have searched for, or other combinations that the user might not otherwise have thought of. In summary, curation, recommendation, and discovery systems offer personalisation of services across millions of users and allow users to find relevant content among the ocean of information online, including not just "more of the same" but also new content they might be interested in (McKelvey & Hunt, 2019).

2.2. Content Moderation

Major social media platforms use filtering mechanisms to identify problematic content and remove or hide it either automatically or after human review. As well as deleting child abuse imagery, terrorist propaganda, copyright infringement, and other illegal content as mandated by national laws, platforms have developed their own community guidelines, enforced by a combination of algorithms and human content moderators. Such rules affect content such as nudity, bullying and harassment, toxic language and hate speech, spam, and deceptive or "fake" accounts (Gillespie, 2018; Saurwein & Spencer-Smith, 2019). Due to the sheer volume of content, this task would be impossible with human labour alone, so platforms use algorithms to deal with this problem of scale (Gillespie, 2018). Pattern-matching content moderation algorithms identify patterns in text, images, video, audio, and user behaviour. These algorithms are continually updated with new information and indicators, known as classifiers, and retrained, so variables and results vary continually. At a high certainty, the algorithms might delete content automatically, and at a lower certainty, the content is sent to a human content moderator (Bradford et al., 2019). In cases where illegal content such as child sexual abuse imagery, terrorist propaganda, and copyright violations have already been identified, the content is provided with a unique identifier called a hash, and can be automatically identified and blocked if users attempt to re-upload it to platforms (Gorwa et al., 2020). While hotly debated due to potential consequences for freedom of expression, the use of algorithms in content moderation enables platforms to quickly remove the most abhorrent kinds of content and helps provide a safer environment for users.

2.3. Allocation of Advertising

As social media platforms do not charge fees to their users, targeted advertising plays a central role in their business models. In contrast to television or print advertising, in which advertisers choose the context in which their advertising is shown, on social media platforms, advertisers can directly select the audience. Social media platforms are able to offer detailed target group definition due to the large quantities of data they hold about users (Busch, 2016). Data is gathered from users' profile information, user behaviour, and their connections. In the US, Facebook previously embellished this with household income and financial data from third party data brokers, although it has since discontinued this practice (Williams & Gebhart, 2018). Platforms are also known to make algorithmic inferences about users from existing data to create new advertising categories. For example, according to information provided by a Facebook spokesperson, "multicultural affinity" is not a category that users assign themselves but is automatically inferred according to pages and posts users have engaged with (Angwin & Parris, 2016). For advertisers, targeted advertising has several advantages over traditional advertising, in terms of automation, accuracy, efficiency, and control. Digital technologies offer more data points to profile individual consumers and allow advertisers to target audiences more precisely. Better profiling and targeting are intended to provide consumers with more relevant information, with which they are more likely to engage (Bodó et al., 2017). Digital formats give advertisers better feedback and control of the process, and allow for experimentation at comparatively lower costs. However, the extensive data collection and processing involved has given rise to concerns about the development of "surveillance capitalism" that comes at the expense of user privacy (Zuboff, 2019).

3. Areas of Algorithmic Harm

While the use of algorithms on social media platforms provides several benefits in terms of user experience and business optimization, it is also accompanied by harms that are subject to increasing public concern. This

section explores these harms and develops a typology that distinguishes harms according to the role that algorithms play in their emergence. In this analysis, evidence of harms involving social media algorithms was collected from media reports and academic papers within a two-year timeframe (from 2018 to 2019), covering Facebook, YouTube, Instagram, and Twitter. The reports were collected through internet searches for the term "social media algorithms" and related search terms in the English and German languages. In a first step, the reports were screened for descriptions of harms and mentions of algorithms in association with harms. As soon as unfamiliar harms were identified, these were investigated further through literature research. In a second step, the harms identified were analysed regarding the causal mechanism of their emergence and the particular role played by algorithms. Harms with similar casual mechanisms were grouped together to inductively develop a typology of algorithmic harms. For example, one kind of harm was caused by users consciously manipulating algorithms with malicious intent. All instances of harm of this nature across different platforms were thus grouped in this harm area. The harm areas were developed inductively until all the instances of harm identified in the study could be assigned to a group. This procedure led to a differentiation of five areas of algorithmic harm:

1. Errors: Algorithms make unsuitable, undesirable or disturbing selections.
2. Manipulation: Algorithms are manipulated by users to produce algorithmic outputs that harass other users or disrupt public discourse.
3. Reinforcement effects: Algorithms strengthen pre-existing harms and inequalities in society.
4. Enablement of harmful practices: Algorithms provide the infrastructure that enables harmful behaviour, e.g., targeted advertising that is opaque and discriminatory.
5. Platform power: Algorithms establish or strengthen platform power over users, markets, and society, and thus pose a challenge to competition, consumers, and individual rights.

3.1. Errors

The first category of algorithmic harm is errors and unsuitable selections by algorithms. Here, algorithms can be seen as the "wrong tool for the job" that make selections lacking human judgement and sensitivity. From a technical standpoint, an algorithm cannot "make a mistake": When a content moderation algorithm deletes a photo of a nude statue, it is carrying out its programmed instructions. From the standpoint of the platform's policy, however, this action is erroneous, because photos of nude statues are not banned. Thus, an algorithmic error refers to an algorithmic decision that produces an outcome at odds with rules, policy, or intention of the algorithm's proprietor.

A well-known example of algorithmic error is the problem of "overblocking" in content moderation. As algorithms are unable to understand the context of a post, this can lead to content being flagged and/or removed when it should not have been. At various points, algorithms have mistaken nudity in art for pornography because they are able to detect patterns that indicate nudity but cannot differentiate between the contexts of art and pornography (Gillespie, 2018). Similarly, algorithms may identify certain keywords or speech patterns as hate speech without being able to evaluate context or intent (Gorwa et al., 2020). As well as impacting freedom of expression, algorithmic errors can have consequences for economic rights, particularly in automated copyright enforcement (Lessig, 1999; Rugnetta, 2018) and on platforms such as YouTube, where users can be sanctioned by losing the ability to earn income from their content (Caplan & Gillespie, 2020).

Alongside enforcement errors, inappropriate recommendations can also be considered as a form of algorithmic selection at odds with the intentions of the platform. Dubbed by Bucher (2016) as "cruel connections," a well-known example of this occurred on Facebook when a user was automatically shown an algorithmically-generated "year in review" album of his posts, which featured a picture of his recently deceased child (Meyer, 2014). Such examples underscore algorithms' lack of understanding for context that a human would have. To summarise, algorithms make unsuitable selections lacking human judgement and sensitivity. However, this harm does not arise from algorithms alone; it can rather be considered an assemblage that encompasses component parts including data that is imbued with social meaning and context, and platforms that seek to automate potentially sensitive tasks using such data.

3.2. Manipulation

Algorithmic selections can be manipulated by users for commercial or abusive purposes, with the outcome that they harass others, disrupt public discourse, and cause harm. This can be seen as a form of "manipulation of institutions or systems," which has at its goal the attainment of covert influence over the people using the platform (Susser et al., 2019, p. 13). Computational propaganda, for instance, refers to "the ways in which the use of algorithms, automation (most often in the form of political bots), and human curation are used over social media to purposefully distribute misleading information" (Woolley, 2020, p. 90). The Russian troll factory Internet Research Agency, for example, allegedly used bots to like and share social media posts from certain accounts so that social media algorithms would consider them popular and be more likely to share them (Osipova & Byrd, 2017). As well as using bots to make content appear more popular than it really is, groups of users can act together in coordinated campaigns to make their content more likely to be recommended by algorithms (Gillespie,

2014). For example, the German far-right internet hate group Reconquista Germanica coordinated their members to post the same hashtags on Twitter at the same time, so that the hashtags would be selected by the algorithm to appear in the top Twitter trends (Kreißel et al., 2018).

It is important to note that the line between "genuine'' behaviours, "legitimate optimization," and "illegitimate, manipulative" behaviours that "game the algorithm" is a thin one (Gillespie, 2017), particularly as users pursuing such strategies, to whatever end, are all ultimately incentivised by the algorithmic logics of the platforms. Indeed, users engage in behaviours designed to manipulate social media algorithms without intending or causing harm to others. On Instagram, users form "engagement pods" in a mutually beneficial arrangement to boost each other's content algorithmically (O'Meara, 2020). However, attempts to exploit the algorithmic logics of platforms can also lead to grotesque consequences, as is evidenced by the "Elsagate" controversy on YouTube, in which inappropriate content, e.g., showing popular children's characters in disturbing situations, are recommended to young audiences on YouTube, who are too young to enter search terms and are thus wholly reliant on recommender systems (Jaakola, 2019). This has resulted in YouTube channels creating increasingly bizarre and troubling content for children by orienting themselves towards the algorithm for commercial purposes (Bridle, 2017). To summarise this kind of harm, the role of the algorithm is as a means that can be manipulated to produce a harmful outcome. This harm does not emerge from the algorithm alone, but from an assemblage that encompasses platforms that offer content recommendations, and thus promise publicity or commercial gain, and users who employ tactics to exploit the logics of algorithms (O'Meara, 2020).

3.3. Reinforcement Effects

Algorithms reinforce, strengthen, or amplify pre-existing phenomena that pose a threat to public discourse and democracy, such as spreading hate speech and disinformation, and entrenching polarisation and radicalisation. Here, algorithms act as a strengthener of, or catalyst for, pre-existing harms that have been present in communication since pre-internet times, but that have been accelerated by the introduction of algorithms. One example of this is the amplification of hate speech and disinformation online. Especially posts that generate strong emotions attract high levels of engagement, which signals high relevance to recommendation algorithms and leads to further recommendation to other users (e.g., Stark et al., 2020, p. 40). Observers claim that algorithms such as the Facebook News Feed algorithm play a role in how hate speech posts go viral, inspiring real-life violence in Sri Lanka (Taub & Fisher, 2018a) and Myanmar (McLaughlin, 2018). When it comes to disinformation, it has been hypothesised that disinformation content achieves amplification by provoking curiosity through novelty, as well as anger through outrage (Vosoughi et al., 2018). In one example, shortly after Facebook shifted its "trending topics" feature from human to algorithmic curation, a number of disinformation stories appeared on it, including a fake story about US journalist Megyn Kelly being fired from Fox News, as the algorithms boosted popular stories without being able to sift out false information (Ohlheiser, 2016).

Another facet of reinforcement is the concept of the "filter bubble" (Pariser, 2011), a personalised media environment that develops when algorithms select content personally tailored to user preferences. While amplification is a phenomenon that occurs across a platform, the filter bubble is generated at the level of the individual user, as algorithms recommend content that fit the algorithmically-assigned interests of that user and the user's activity on the platform provides further feedback to the algorithm. It is argued that algorithms reinforce confirmation bias because they predominantly deliver opinions that affirm pre-existing beliefs and mislead users into believing that everyone else holds the same opinions as them, creating an echo chamber. Echo chambers can also induce people to believe that hatred of a particular group is the social norm (Taub & Fisher, 2018b). It is hypothesised that algorithmic personalisation reduces exposure to different content and new ideas, with potentially negative outcomes for innovation and the development of new ideas (Sunstein, 2001). However, empirical studies have suggested that the impact of filter bubbles is limited (Zuiderveen Borgesius et al., 2016) and moderated by an environment in which a variety of different media continue to be consumed (Dubois & Blank, 2018). Recently, however, the phenomenon of "rabbit holes" has come to attention in the media, in which recommendation algorithms contribute to radicalisation of users by recommending more and more extreme content, such as conspiracy theories (Lewis, 2018). Furthermore, research suggests that algorithmically-mediated advertising on social media reinforces gender and age stereotyping by showing ads to users that fit stereotypes, such as showing advertising about beauty to women or fashion to younger people (Bol et al., 2020), and that not just advertiser choice or user preferences play a role, but also algorithmic selection (Ali et al., 2019). By charging more per click for advertising to audiences that are not in the perceived core market for an ad, platforms may also be making it more difficult for political parties to break through the "filter bubble" to reach users outside their traditional voter base (Ali et al., 2021).

To summarise, the role of the algorithm is as a technology that reinforces problematic content and harmful conduct. Here, algorithms are part of the interplay between content, platform logics and user behaviour. In particular, the algorithms in question operate in the context of recommender systems and are thus engineered to recommend content with high levels of

engagement. The harm arises when it promotes content that has little impact on society in small quantities but becomes problematic when it is amplified across many users or reinforces problematic worldviews in individual users (Cobbe & Singh, 2019). In addition, a role is played by the large numbers of users who engage with such content by liking, commenting, sharing, and clicking on it.

3.4. Enabling Harmful Practices

Algorithms can also enable actors to carry out discriminatory practices, particularly through online advertising. Here, algorithms are used as infrastructure to target or exclude certain groups of users, with harmful effects. For example, Facebook uses data and algorithms to determine if users in the US belong to an ethnic minority for the purpose of advertising to those ethnic groups. However, the same functions have been used for manipulative purposes. The Trump 2016 presidential campaign disclosed that it targeted Facebook ads to African Americans to discourage them from voting (Green & Issenberg, 2016).

The same functions also made it possible to exclude ethnic minorities from seeing certain ads, as journalists have found instances where it was possible to exclude ethnic minorities from seeing ads for housing and accommodation (Angwin et al., 2017; Cotter et al., 2021). Facebook disabled advertisers' ability to exclude ethnic minorities at the end of 2017, but the incident nonetheless shows how platforms have not carefully considered how automated, targeted advertising can be used to suppress and discriminate against marginalised groups. In addition, targeted advertising ensures that it is only seen by the target audience and not by others. This is particularly troubling in political microtargeting, as a political advertiser can send different voters different, contradictory information while avoiding broader public scrutiny. This decreases the transparency of campaigns, political positions, and electoral promises and could lead to a skewed perception of priorities of political parties among voters (Zuiderveen Borgesius et al., 2018, pp. 87–89). To summarise, the role of the algorithm is as an infrastructure that enables harmful practices, such as discrimination. That said, not the algorithm alone is at fault: It is rather part of an assemblage of the infrastructure of online advertising that is intentionally designed to include and exclude segments of the audience to optimize targeting, as well as the social ills that can be strengthened by such techniques.

3.5. Platform Power

Algorithms may strengthen platform power, particularly over competitors, markets, and users. Here, the role of algorithms is as a tool of influence and surveillance over other actors. The use of big data and algorithms can enable a "God view," using "big data and big analytics for a clearer overview of the marketplace at any given moment" (Ezrachi & Stucke, 2016, p. 72). For example, the Facebook app Onavo Protect offered users a VPN service while also collecting data on how users use competitor apps. The information is alleged to have informed Facebook's decisions about which app features to imitate, including stories from Snapchat, and which companies to acquire, including WhatsApp and Instagram (Seetharaman & Morris, 2017). Monitoring through Onavo Protect may be one of a number of anticompetitive practices in the technology market ("American tech giants," 2018).

Algorithms have also contributed to the unequal relationship between platforms and certain markets, particularly in fields of publishing that are particularly dependent on social media platforms for distribution. The rise of algorithm-based targeted advertising on internet platforms has contributed to the disruption of traditional funding models for journalism (Lobigs, 2016, pp. 103–104), and publishers have become increasingly dependent on social media to the extent that Facebook has been described as a "kingmaker" (Pasquale, 2015). The dependence of publishers on social media algorithms is exemplified by Facebook's shifting priorities when it comes to video. When Facebook increased the importance of video content in the News Feed algorithm in 2015, publishers active on social media responded by moving resources to video production (Griffith, 2015). This, however, proved short-lived as Facebook decided to assign less priority to video in the algorithm three years later (Vogelstein, 2018) and the same publishers then made social video employees redundant (Bilton, 2018). The pivot to video can be seen as an example of how algorithms are used to impose the changing commercial interests of a social platform on sectors of the publishing industry that are particularly vulnerable to algorithmic change (Oremus, 2018). Indeed, social media creators who are commercially active on platforms are particularly impacted by changes in recommendation and content moderation algorithms, leading to "algorithmic precarity" (Duffy, 2020). Social media creators carry a higher level of exposure to algorithmic change, and thus experience heightened algorithmic precarity, due to their particular dependence on platforms for distribution.

Finally, algorithms strengthen platform power over users by promoting addictive behaviour and eroding privacy. Although more empirical research is needed, it is believed, for example, that social media platforms use algorithms to withhold and distribute likes and notifications so that users keep checking the app (Peitz, 2017). The use of algorithms also raises complex questions about personal privacy and informational self-determination, especially regarding the use of inferential analytics, in which algorithms make inferences about users, often without their consent (Wachter & Mittelstadt, 2019). A further concern is facial recognition technology (Wolfangel, 2018). The application for a patent for the use of facial recognition for payments

(Moore Davis, 2016), as well as a patent for eye-tracking technology (San Agustin Lopez et al., 2014), both by Facebook, has generated speculation about the depth of observation and data gathering that users may be subjected to in the future. In addition, the US news website *The Intercept* claims to have seen a confidential document in which Facebook outlines a new advertising service that will be able to predict users' future consumer behaviour using machine learning (Biddle, 2018). Such applications fuel fears about the potential for surveillance and social scoring, as well as about consumers' continuing ability to make purchasing choices without covert psychological influence.

To summarise, algorithms are used to strengthen influence and surveillance over other actors, and increase platform power over competitors, markets, and users. The assemblage that produces this harm encompasses commercial platforms upon which user interactions and economic activity take place, all mediated by the respective platform company. The concentration of such power by platforms requires significant quantities of data and algorithms that are able to process them. In turn, these data and algorithms enable platforms to conduct surveillance, as well as to intervene and exert influence to pursue their own goals (Zuboff, 2019).

4. Summary and Conclusion

The aim of this article was to contribute to the understanding of algorithmic harm by exploring the roles that algorithms play in the emergence of harms on social media platforms. The article therefore developed a typology of five areas of algorithmic harm based on the mechanisms of their causation. This analysis demonstrated that algorithms contribute to the emergence of harm in manifold ways. Algorithms can be deficient tools that lead to errors, instruments that serve manipulation, technologies that reinforce and amplify problematic content, enabling infrastructure for problematic behaviour and instruments that serve to establish or strengthen platform power.

However, the analysis also found that harms do not arise from the application of algorithms alone. Instead, harms can be best conceived of as socio-technical assemblages that encompass the use and design of algorithms, platform design, commercial interests, social practices, and context. Altogether, these findings support the suggestion that algorithms are not isolated technical artefacts, but "assemblages of institutionally situated code, human practices and normative logics" (Ananny, 2016, p. 108). It is thus useful to understand how they "work within socio-technical assemblages and how they perform actions and make a difference in particular domains" (Kitchin, 2017, p. 26). This is particularly evident on social media platforms, where algorithms and their implications are inseparable from platform architectures, normative logics, and commercial interests of platform companies (van Dijck, 2013).

In addition, it should be considered that types of harm are not isolated from one another but can interact and intersect. Boundaries between types of harm are porous and permeable. For example, users manipulate content moderation algorithms to produce errors, in an event in which algorithms are both deficient tools that lead to errors and instruments of manipulation. After the Christchurch terror attacks in 2019, social media platforms struggled to prevent users from uploading footage filmed by the shooter, in part because users employed techniques designed to bypass content moderation algorithms, such as superimposing footage of YouTube personalities to make the upload look like video game footage (Timberg et al., 2019). A further technique was to upload the footage as a live stream, preventing the video from being analysed by content moderation algorithms as a fully uploaded file (McDonald, 2019). User evasion of content moderation algorithms is an example of interlocking algorithmic harms that may provide further avenues for research.

The analysis of algorithmic harms inevitably leads to questions as to whether and how to deal with them. From an institutional perspective, options range from market solutions and users' own strategies for countering harms, via voluntary industry self-regulation to command-and-control regulation by state authorities (Latzer et al., 2006). Some algorithmic harm could be reduced by consumers' self-help strategies (opting out of services, switching to other providers and technical self-protection, such as privacy tools; Saurwein et al., 2015). However, there are several barriers to effective self-help, and the potential of user self-protection should not be overestimated. Users may not be able to avoid using services or switch to other providers because of network effects and other barriers. Privacy tools may be able to limit the use of cookies, but do not prevent platforms from gathering data on user behaviour on their services. Moreover, because of the opaque nature of algorithmic selection and low levels of awareness about algorithms among users, algorithmic harm is often barely noticeable to consumers. For example, an average internet user can hardly detect errors, reinforcement of problematic phenomena, or manipulation. Consequently, it is argued that if harms and risks are not visible, then there is no reason to consider self-protection strategies (Saurwein et al., 2015). In practice, however, some countries (e.g., Switzerland) report a considerable level of awareness of algorithms and algorithmic harms (Latzer et al., 2020) while in other countries (e.g., Norway) awareness of algorithms is rather low (Gran et al., 2020). For Germany, Fischer and Petersen (2018) report a widespread unawareness of algorithms, strong indecision about risks and opportunities, discomfort over algorithmic decision-making, and a strong desire for more control.

Regarding control, the typology of harms allows us to reflect upon suitable governance responses (Latzer et al., 2019) by exploring incentives for social media

platforms to reduce harms by means of platform self-regulation. Industry self-regulation is unsuitable in cases where harms are also indistinguishable from the commercial interests of industry players. This is particularly evident in growing platform power, where social media platforms have the least motivation to reduce algorithmic harms. Thus, the current focus of statutory regulation on data protection and antitrust is well justified. In the case of algorithmic errors, however, there are clear incentives for platforms to reduce errors and increase accuracy because functioning automation is key to the further scaling up of services. In content moderation, for instance, platforms have been making efforts to improve the accuracy of automated moderation systems and regularly report performance indicators to demonstrate progress. In the case of manipulation by third parties there are some incentives for platform providers to combat manipulation and maintain the integrity and reputation of their services. Platforms make use of their terms of service to define unwanted behaviour and have made efforts to identify and sanction inauthentic behaviour and block bot accounts. A continual challenge is drawing the line between legitimate optimization and illegitimate gaming.

Compared to errors and manipulation, incentives to counter the reinforcement of problematic phenomena are less clear-cut. It can be argued that amplification of problematic content contributes to profitability, which reduces incentives to curb it. On the other hand, platforms may be motivated to control amplification when it starts to impair user experience and discourage users from spending time on the platforms. Indeed, Facebook now deprioritises "borderline content" in its News Feed algorithm (Zuckerberg, 2018) and the major platforms have proven more willing to act against the spread of problematic content in the context of the Covid-19 pandemic and threats to US democracy that culminated in the storming of the Capitol in January 2021. Similarly, when it comes to harmful advertising practices, platforms have been motivated to disable some problematic features and improve transparency through a political advertising database only after the issue became a public relations problem. This is illustrative of a broader pattern of platform governance as a cycle of "shocks and exceptions" (Gillespie & Ananny, 2016).

Moreover, platforms may be more or less motivated to address harms depending on who is affected by them. A reliance on public relations shocks means a reliance on journalism as a mechanism for uncovering algorithmic harms, as well as other online harms (Diakopoulos, 2015). Considering that most users do not have access to this kind of publicity, relying on journalism as a principal accountability mechanism is not a sustainable means of reducing harm. Furthermore, bias and insufficient employee diversity within platform companies create a blind spot towards algorithmic harms that affect groups who are commonly discriminated against and marginalised in society (Benjamin, 2019; Noble, 2018). These factors could slow the response of companies in addressing harms that affect users who do not have access to publicity or are structurally oppressed in society. The failures of social media companies in addressing algorithmic harm have led to a growing call to increase statutory regulation and oversight. Most recently, the European Commission published a legislative initiative for a Digital Services Act to enhance platform accountability (European Commission, 2020). The proposed regulations also concern algorithm-based services such as recommendation systems, content moderation, and advertising. The regulations shall force very large online platforms to increase the transparency of their algorithmic systems, to provide opportunities to opt-out from profiling and personalisation, to protect services from manipulation, and to carry out risk assessments to avoid the spread of illegal content, restrictions of fundamental rights, and manipulation. The proposal suggests the establishment of external and independent auditing procedures and "technical assistance at EU level, for inspecting and auditing content moderation systems, recommender systems and online advertising" (European Commission, 2020, p. 12). The discussion of a Digital Services Act is at an early stage, but the legislative initiative clearly indicates that algorithmic harms have become a prominent issue on the internet governance agenda, which may lead to stronger control of internet platforms and their algorithm-based modes of operation.

The limitations of our study regarding its scope provide potential impulses for future research. While the article has analysed algorithms on social media platforms, further research could investigate algorithmic harm across other kinds of platforms, such as Amazon and Uber, building upon existing critiques of individual platforms (see Khan, 2017; Muller, 2020). Furthermore, this research focused on platforms popular in North America and Europe, and used sources in the English and German languages, limiting its geographical and cultural scope. Future avenues of research could include investigations of algorithmic harm across non-Western cultural contexts, in particular in areas such as algorithmic content moderation on large global platforms, where implementation across languages and geographic regions is uneven. Finally, analyses of algorithmic harms lead to questions of suitable governance responses. The article provides a set of theoretical reflections upon the incentives for social media platforms to reduce harms by means of platform self-regulation. Future research should verify if the governance of algorithms in fact coincides with the proposed patterns.

Acknowledgments

The article presents results from the project "The Automation of the Social: Algorithmic Selection in Social Online Networks" funded by the Vienna Anniversary Fund for the Austrian Academy of Sciences.

Conflict of Interests

The authors declare no conflict of interests.

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*. https://doi.org/10.1145/3359301

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2021). Ad delivery algorithms: The hidden arbiters of political messaging. In *WSDM '21: Proceedings of the 14th ACM international conference on web search and data mining* (pp. 13–21). Association for Computing Machinery.

American tech giants are making life tough for startups. (2018, June 2). *The Economist*. https://www.economist.com/business/2018/06/02/american-tech-giants-are-making-life-tough-for-startups

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117.

Angwin, J., & Parris, T. (2016, October 28). Facebook lets advertisers exclude users by race. *ProPublica*. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

Angwin, J., Tobin, A., & Varner, M. (2017, November 21). Facebook (still) letting housing advertisers exclude users by race. *ProPublica*. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity Press.

Bergen, M. (2019, April 2). YouTube executives ignored warnings, letting toxic videos run rampant. *Bloomberg*. https://www.bloomberg.com/news/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

Biddle, S. (2018, April 13). Facebook uses artificial intelligence to predict your future actions for advertisers, says confidential document. *The Intercept*. https://theintercept.com/2018/04/13/facebook-advertising-data-artificial-intelligence-ai

Bilton, R. (2018, February 21). Post-Facebook News Feed tweaks, Vox Media lays off 50 employees. *Nieman Lab*. http://www.niemanlab.org/2018/02/post-facebook-news-feed-tweaks-vox-media-lays-off-50-employees

Bodó, B., Helberger, N., & de Vreese, C. H. (2017). Political micro-targeting: a Manchurian candidate or just a dark horse? *Internet Policy Review*, *6*(4), 1–13.

Bol, N., Strycharz, J., Helberger, N., van de Velde, B., & de Vreese, C. H. (2020). Vulnerability in a tracked society: Combining tracking and survey data to understand who gets targeted with what content. *New Media & Society*, *22*(11), 1996–2017.

Bradford, B., Grisel, F., Meares, T., Owens, E., Pineda, B., Shapiro, J., Tyler, T., & Evans Peterman, D. (2019). *Report of the Facebook data transparency advisory group.* Yale Law School. https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf

Bridle, J. (2017). *Something is wrong on the internet*. Medium. https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

Bucher, T. (2016). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44.

Busch, O. (2016). The programmatic advertising principle. In O. Busch (Ed.), *Programmatic advertising. The successful transformation to automated, data-driven marketing in real-time* (pp. 3–15). Springer.

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 1–13.

Cobbe, J., & Singh, J. (2019). Regulating recommending: Motivations, considerations, and principles. *European Journal of Law and Technology*, *10*(3), 1–37.

Cotter, K., Medeiros, M., Pak, C., & Thorson, K. (2021). "Reach the right people": The politics of "interests" in Facebook's classification system for ad targeting. *Big Data & Society*, *8*(1), 1–16.

DeVito, M. (2017). From editors to algorithms. *Digital Journalism*, *5*(6), 753–773.

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

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745.

Duffy, E. B. (2020). Algorithmic precarity in cultural work. *Communication and the Public*, 5(3/4), 103–107.

European Commission. (2020). *Proposal for a regulation of the European Parliament and the Council on a single market for digital services (Digital Services Act) and amending directive 2000/31/EC* (COM(2020)825). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN

Ezrachi, A., & Stucke, M. (2016). *Virtual competition*. Harvard University Press.

Facebook Investor Relations. (2021). *Facebook reports fourth quarter and full year 2020 results*. https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx

Fischer, S., & Petersen, T. (2018*). Was Deutschland über Algorithmen weiß und denkt. Ergebnisse einer repräsentativen Bevölkerungsumfrage* [What Germany knows and thinks about algorithms. Results from a representative population survey]. Gütersloh.

Gillespie, T. (2014). The relevance of algorithms. In T.

Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–194). MIT Press.

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society*, *20*(1), 63–80.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T., & Ananny, M. (2016). Public platforms: Beyond the cycle of shocks and exceptions. *Oxford Internet Institute*. Retrieved from http://blogs.oii.ox.ac.uk/ipp-conference/2016/programme-2016/track-b-governance/platform-studies/tarleton-gillespie-mike-ananny.html

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15.

Gran, A.-B., Booth, P., & Bucher, T. (2020). To be or not to be algorithm aware: a question of a new digital divide? *Information, Communication & Society*. Advance online publication. https://doi.org/10.1080/1369118X.2020.1736124

Green, J., & Issenberg, S. (2016, October 26). Why the Trump machine is built to last beyond the election. *Bloomberg*. https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go

Griffith, E. (2015, June 3). How Facebook's video-traffic explosion is shaking up the advertising world. *Fortune*. http://fortune.com/2015/06/03/facebook-video-traffic

Hale, J. (2019, May 7). More than 500 hours of content are now being uploaded to YouTube every minute. *Tubefilter*. https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute

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

Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society*, *1*(2), 1–11.

Jaakola, M. (2019). From vernacularized commercialism to kidbait: Toy review videos on YouTube and the problematics of the mash-up genre. *Journal of Children and Media*, *14*(2), 237–254.

Khan, L. M. (2017). Amazon's antitrust paradox. *Yale Law Journal*, *126*(3), 710–805.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29.

Kreißel, P., Ebner, J., Urban, A., & Jakob, G. (2018). *Hass auf Knopdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz* [Hate at the touch of a button. Far-right troll factories and the ecosystem of coordinated hate campaigns on the Internet]. Institute for Strategic Dialogue.

Latzer, M., Festic, N., & Kappeler, K. (2020). *Awareness of risks related to algorithmic selection in Switzerland.* University of Zurich. https://mediachange.ch/media//pdf/publications/Report_3_Risks.pdf

Latzer, M., Hollnbuchner, K., Just, N., & Saurwein, F. (2016). The economics of algorithmic selection on the Internet. In J. Bauer & M. Latzer (Eds.), *Handbook on the economics of the internet* (pp. 395–425). Edward Elgar.

Latzer, M., Just, N., Saurwein, F., & Slominski, P. (2006). Institutional variety in communications regulation. Classification scheme and empirical evidence from Austria. *Telecommunications Policy*, *30*(3/4), 152–170.

Latzer, M., Saurwein, F., & Just, N. (2019). Assessing policy II: Governance-choice method. In H. van den Bulck, M. Puppis, K. Donders, & L. van Audenhove (Eds.), *The Palgrave handbook of methods for media policy research* (pp. 557–574). Palgrave Macmillan.

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

Lewis, P. (2018, February 2). 'Fiction is outperforming reality': How YouTube's algorithm distorts truth. *The Guardian*. https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth

Lobigs, F. (2016). Finanzierung des Journalismus—Von langsamen und schnellen Disruptionen [Financing of journalism—Of slow and fast disruptions]. In K. Meier & C. Neuberger, C. (Eds.), *Journalismusforschung. Stand und Perspektiven* [Journalism research. Current state and perspectives] (pp. 69–137). Nomos.

McDonald, S. (2019, March 15). Google AI has trouble keeping NZ massacre video off YouTube. *Newsweek*. https://www.newsweek.com/google-ai-has-trouble-keeping-nz-massacre-video-youtube-1365375

McKelvey, F., & Hunt, R. (2019). Discoverability: Toward a definition of content discovery through platforms. *Social Media + Society*, *5*(1), 1–15.

McLaughlin, T. (2018, July 6). How Facebook's rise fuelled chaos and confusion in Myanmar. *Wired*. https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar

Meyer, E. (2014). *Inadvertent algorithmic cruelty*. Meyerweb. http://meyerweb.com/eric/thoughts/2014/12/24/inadvertent-algorithmic-cruelty

Moore Davis, S. (2016). *Facial recognition identification for in-store payment transactions* (US Patent No. US20170323299). Patent and Trademark Office.

Muller, Z. (2020). Algorithmic harms to workers in the platform economy: The case of Uber. *Columbia Journal of Law and Social Problems*, *53*(2), 167–210.

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

Ohlheiser, A. (2016, August 29). Three days after removing human editors, Facebook is already trending fake news. *Washington Post*. https://www.washingtonpost.com/news/the-intersect/wp/2016/08/29/a-fake-headline-about-megyn-kelly-was-trending-on-facebook

O'Meara, V. (2020). Weapons of the chic: Instagram influencer engagement pods as practices of resistance to Instagram platform labor. *Social Media + Society*, *5*(4), 1–11.

Oremus, W. (2018, October 18). The big lie behind the 'pivot to video.' *Slate Magazine*. https://slate.com/technology/2018/10/facebook-online-video-pivot-metrics-false.html

Osipova, N. V., & Byrd, A. (2017, October 31). Inside Russia's network of bots and trolls. *New York Times*. https://www.nytimes.com/video/us/politics/100000005414346/how-russian-bots-and-trolls-invade-our-lives-and-elections.html

Pariser, E. (2011). *The filter bubble*. Penguin Press.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Peitz, D. (2017, July 17). Erstmals geben Tech-Leute zu: Wir haben ein echtes Problem [Tech people admit for the first time: We have a real problem]. *Die Zeit*. https://www.zeit.de/digital/2018-07/smartphone nutzung-sucht-david-levy-computerwissenschaftler

Persily, N., & Tucker, J. A. (Eds.). (2020). *Social media and democracy. The state of the field, prospects for reform*. Cambridge University Press.

Rugnetta, M. (2018). *Automated copywrongs*. Reasonably Sound. http://reasonablysound.com/2018/01/15/automated-copywrongs

San Agustin Lopez, J., Sztuk, S., & Henrik Tall, M. (2014). *Systems and methods of eye tracking control* (US Patent No. US9829971B2). Patent and Trademark Office. https://patents.google.com/patent/US9829971B2/en

Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: Options and limitations. *Info: The Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media*, *17*(6), 35–49.

Saurwein, F., & Spencer-Smith, C. (2019). *Inhaltsregulierung auf Internet-Plattformen. Optionen für verantwortungsvolle Governance auf nationaler Ebene* [Content moderation on internet platforms. Options for accountability-oriented governance at national level] (Research Report). CMC.

Seetharaman, D., & Morris, B. (2017, August 13). Facebook's Onavo gives social-media firm inside peek at rivals' users. *Wall Street Journal*. https://www.wsj.com/articles/facebooks-onavo-gives-social-media-firm-inside-peek-at-rivals-users-1502622003

Stark, B., Stegmann, D., Magin, M., & Jürgens, P. (2020). *Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse*. AlgorithmWatch.

Sunstein, C. (2001). *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, *4*(1), 1–45.

Taub, A., & Fisher, M. (2018a, April 21). Where countries are tinderboxes and Facebook is a match. *New York Times*. https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html

Taub, A., & Fisher, M. (2018b, August 21). Facebook fueled anti-refugee attacks in Germany, new research suggests. *New York Times*. https://www.nytimes.com/2018/08/21/world/europe/facebook-refugee-attacks-germany.html

Timberg, C., Harwell, D., Shaban, H., Ba Tran, A., & Fung, B. (2019, March 15). The New Zealand shooting shows how YouTube and Facebook spread hate and violent images—yet again. *Washington Post*. https://www.washingtonpost.com/technology/2019/03/15/facebook-youtube-twitter-amplified-video-christchurch-mosque-shooting

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colorado Technology Law Journal*, *13*(2), 203–218.

US House Judiciary Subcommittee on Antitrust, Commercial, and Administrative Law. (2020). *Investigation of competition in digital markets*.

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society*. Oxford University Press.

Vogelstein, F. (2018, January 13). Facebook's Adam Mosseri on why you'll see less video, more from friends. *Wired*. https://www.wired.com/story/facebooks-adam-mosseri-on-why-youll-see-less-video-more-from-friends

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, *2019*(2), 494–620.

Williams, J., & Gebhart, G. (2018). *Facebook isn't telling the whole story about its decision to stop partnering with data brokers.* Electronic Frontier Foundation. https://www.eff.org/de/deeplinks/2018/04/facebook-isnt-telling-whole-story-about-its-decision-stop-partnering-data-brokers

Wolfangel, E. (2018, March 5). Facebook: Gesichtserkennung lässt sich eben nicht abschalten [Facebook: Facial recognition cannot be turned off]. *Spektrum*. https://www.spektrum.de/kolumne/gesichtserkennung-laesst-sich-eben-nicht-abschalten/1548879

Woolley, S. C. (2020). Bots and computational propaganda: Automation for communication and control.

In N. Persily & J. A. Tucker (Eds.), *Social media and democracy. The state of the field, prospects for reform* (pp. 89–110). Cambridge University Press.

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

Zuckerberg, M. (2018). *A blueprint for content governance and enforcement*. Facebook. https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634

Zuiderveen Borgesius, F. J., Moeller, J., Kruikemeier, S., Fathaigh, R., Irion, K., Dobber, T., Bodó, B., & de Vreese, C. H. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, *14*(1), 82–89.

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., de Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, *5*(1), 1–16.

About the Authors

Florian Saurwein is a senior scientist at the Institute for Comparative Media and Communication Studies (CMC) of the Austrian Academy of Sciences and the University of Klagenfurt. He studied communication science and political science, and holds a PhD from the University of Zurich. His research centres around interrelations of media, society, and governance, with a current focus on risks and governance of content moderation and algorithmic selection on internet platforms.

Charlotte Spencer-Smith is a doctoral candidate at the Department of Communication Studies of the Paris Lodron University of Salzburg, Austria. She was previously a senior scientist without doctorate at the Institute for Comparative Media and Communication Studies of the Austrian Academy of Sciences and the University of Klagenfurt.