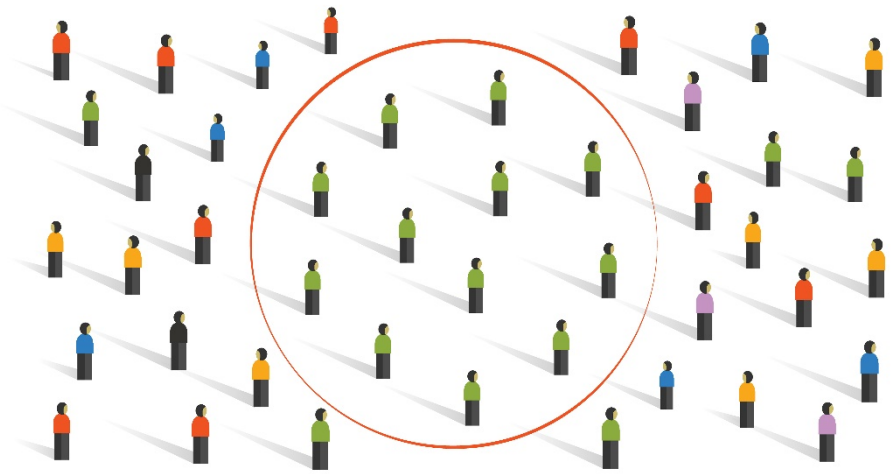
European Parliament

Key social media risks to democracy

Risks from surveillance, personalisation, disinformation, moderation and microtargeting



IN-DEPTH ANALYSIS

EPRS | European Parliamentary Research Service

Author: Costica Dumbrava
Members' Research Service

PE 698.845 – December 2021

EN

Once regarded as great enablers of democracy, social media are nowadays blamed for many of the ailments of democracy. They are criticised for spreading disinformation, sowing discord, manipulating citizens and undermining democratic institutions. Why are social media important for democracy? What are the main risks posed by social media to different dimensions of democracy, such as political participation, electoral processes, and democratic institutions? What is the role of algorithms? To what extent are various concerns about social media backed by empirical evidence? This analysis provides an overview of the main risks posed by social media to democracy, linked to surveillance, personalisation, disinformation, moderation and microtargeting. Furthemore, it discusses key approaches to tackling social media risks to democracy in the context of relevant ongoing EU legislative and policy work.

AUTHOR

Costica Dumbrava, Members' Research Service, EPRS

This paper has been drawn up by the Members' Research Service, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

To contact the authors, please email: eprs@ep.europa.eu

LINGUISTIC VERSIONS

Original: EN

Translations: FR, DE

Manuscript completed in December 2021.

PE 698.845
ISBN: 978-92-846-8802-9
DOI:10.2861/135170
QA-08-21-390-EN-N

eprs@ep.europa.eu
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary

Democracy relies on citizens' abilities to obtain information on public matters, to understand them and to deliberate about them. Whereas social media provide citizens with **new opportunities** to access information, express opinions and participate in democratic processes, they can also **undermine democracy** by distorting information, promoting false stories and facilitating political manipulation. Social media risks to democracy can be classified according to **five aspects** that generate risks surveillance, personalisation, disinformation, moderation, and microtargeting.

Firstly, social media provide new and more effective ways to monitor people online, which can be used by governments to target politically active citizens and silence dissent (**political surveillance**). Even in the absence of explicit coercion, citizens who suspect they are the target of online surveillance may suppress their political expression online for fear of retribution. The massive collection of data by social media creates privacy risks to users and may affect their capacity to form and express political opinions (**loss of privacy and autonomy**). The attention capture model used by social media seeks to exploit human needs and biases in order to increase engagement, but at the same time it undermines individual autonomy. Social media may also contribute to citizens' decreasing levels of interest in politics, even if they are not directly responsible for this (**political disengagement**). Certain effects of social media are a by-product of a particular business model focused on engagement at all costs. This indifference of social media to democracy contrasts with the fact that they have a growing impact on democracy.

Secondly, the promotion of personalised content on social media may lock citizens in informational bubbles, thus affecting their capacity to form opinions (**narrowed worldviews**). Whereas content personalisation can help citizens deal with the problem of information overload, it can also limit the range of information available to them. Moreover, the segmentation of information and engagement may reinforce group boundaries and reduce opportunities for political dialogue (**social and political fragmentation**). Yet, despite widespread concern, existing empirical evidence suggests that the personalisation and filtering effects of social media are less severe and pervasive than initially feared. Whereas the negative political effects of personalisation seem less severe and widespread, the risk of societal fragmentation and polarisation remains. It must be noted that evaluations of the political effects of social media may also depend on political (ideological) assumptions about the nature and conditions of democratic politics.

Thirdly, the spread of false information on social media can undermine citizens' capacity to form and express political views (**distortion of political views and preferences**). Despite growing evidence of people's significant exposure to political disinformation online, the actual impact of disinformation on their views and preferences is difficult to assess. Although the reach and impact of disinformation seem to have been over-estimated, there is evidence of negative effects in particular contexts and on specific groups. Disinformation can be used to persuade or confuse voters and to mobilise or demobilise citizens to cast a vote, which may, in certain conditions, be a determinant of election outcomes (**distortion of electoral outcomes**). Importantly, widespread disinformation and acute public perception thereof (amplified by lack of research and inadequate reporting) may undermine trust in (all) online information and democratic institutions. Despite recent media attention being focused primarily on disinformation disseminated by foreign actors (e.g. foreign governments or intermediaries seeking to influence electoral outcomes in another country), disinformation is also spread by domestic actors (e.g. political parties and politicians seeking to influence pulblic opinion in their own country). Sometimes, this happens as a result of entrepreneurs promoting highly engaging content to make profits from selling ads. Moreover, automated accounts and algorithms contribute to the spread of disinformation on social media (**automated disinformation**). However, effective disinformation campaigns are a result of a complex interaction between humans and algorithms. For example, automated tools for spreading

false information exploit human biases and predispositions, such as human confirmation bias, inclination to believe repeated stories, and attraction to novel content.

Fourthly, efforts by social media platforms to tackle disinformation and other forms of deception online may undermine users' freedom of expression and enable control over public opinion (**political censorship**). Whereas all moderation measures are risky, content removal is particularly problematic when targeted content is not explicitly illegal. Deleting and labelling content can be counterproductive, as it may reinforce perceptions about unfair and unjustified censorship of particular views and groups. Whereas automation can alleviate some burdens of human moderation, it can also amplify errors and automate pre-existing bias (**algorithmic bias**). Increased pressure on social media to tackle problematic content may push platforms to rely even more strongly on automated tools, which leads to more censorship and bias. Despite efforts to make moderation more transparent and systematic, moderation measures adopted by social media remain largely unclear, arbitrary, and inconsistently applied. The risk is that social media platforms take decisions with significant consequences for individuals and democracy without proper accountability (**lack of accountability**).

Fifthly, social media platforms rely on a variety of user data to profile people and sell targeted advertising (microtargeting). Whereas political microtargeting can serve to re-engage citizens in politics, it can also be used to manipulate citizens' views and expectations (**political manipulation**). The covert or hidden nature of microtargeting increases the risk of manipulation and thus undercuts citizens' capacity to form and make political choices. Political microtargeting also challenges existing electoral rules concerning transparency, campaigning and political funding, and can distort elections (**distortion of the electoral process**). Whereas evidence about the widespread use of political microtargeting is growing, its actual impact remains uncertain. Given the nature of political competition, it is possible that political microtargeting campaigns can determine the outcome of elections, in particular in winner-takes-all electoral systems. Even if microtargeting cannot be blamed for tipping recent elections, the risks it creates are likely to increase, given the high political and economic interests at stake and future technological advances.

The EU already has laws and policies in place to tackle many of the social media risks to democracy (for example, strong data protection rules) and is spearheading efforts to counteract new challenges (such as new legislative proposals on digital services). There are seven **key approaches** to tackling social media risks to democracy.

EU competition measures can be used to further combat abuses of market dominance, for example, by controlling social media platforms' ability to integrate behavioural data from various services and advertising networks and by promoting data portability and interoperability solutions to reduce the cost of switching between platforms. Further clarification and stricter enforcement of **EU data protection and digital privacy** rules can help to prevent abuses of personal data and provide safeguards for fair and democratic elections. Amid widespread calls for increasing social media responsibility for promoted content, there is an ongoing reflection on the need to review and clarify **EU content liability** rules on online content. Special attention has also been given to **increasing transparency and accountability of online platforms** for filtering and moderating content, including for the **use of algorithms**. The EU is gradually moving towards a co-regulatory approach that would require social media platforms to assume stricter transparency and accountability obligations. Specific rules are also forthcoming to prevent abuse and manipulation through **targeted political advertising.** Lastly, addressing the social media risks to democracy cannot succeed without **empowering citizens** to understand and fend off online risks, for example, by improving digital literacy, promoting citizen-centred approaches to tackling online challenges, and supporting public-oriented institutions such as independent media.

Table of contents

Table of tables

1. Social media and democracy

Social media platforms have recently come under **heavy criticism** for a variety of reasons, including, among others, for disregarding people's rights to data protection and privacy, facilitating the spread of false information, aiding political manipulation and undermining citizens' freedom of expression. To assess these concerns, it is necessary to understand why democracy is vulnerable to social media, what specific risks this vulnerability entails and, considering existing evidence, to what extent these risks have materialised.

1.1. Informational conditions of democracy

Democracy stands for a **system of government in which it is the people who rule**. In the contemporary context, this means that people have some form of participation in the government and that even if people do not participate directly in all government decisions, the authority of the government is derived from the people. Democracies come in different kinds of shapes and shades. Moreover, there are generally several key dimensions of democracy in terms of fundamental moral principles, basic rights and institutions, and empirical preconditions.

At the heart of the ideal of democracy are the moral **principles of autonomy and equality.** Firstly, autonomy refers to the capacity of individuals to conceptualise, formulate and choose norms for themselves to follow. Democracy recognises this moral capacity as an individual right to political self-determination by including all individuals in the political decision-making process.[1] Political equality means that all individuals are equally competent and should be able to participate equally in the political decision-making. In a representative democracy, the right to political self-determination is most apparent in the equal participation of citizens in the election of political representatives (equality of the vote and universal franchise).

In practice, the moral principles of democracy are realised through a set of **rights and institutions**. Citizens enjoy a series of political rights, such as the right to vote and freedom of association, thought and expression. The law also prescribes rules for the functioning of political institutions and processes (e.g. electoral systems). Modern democracies are constitutional, meaning that citizens enjoy a set of fundamental rights and freedoms regardless of the electoral outcomes or the political orientation of a particular government.

There is a long debate about what makes a particular country or system democratic and what **conditions are needed for democracy** to survive and flourish. In a classic account, American political scientist Robert Dahl posited that a democratic process should meet five necessary conditions.[2] One of these involves 'enlightened understanding', which requires that each member (citizen) should have equal opportunities to learn about the relevant alternative policies and their likely consequences. In a more recent contribution, Cohen and Fung identified several informational conditions for a democratic public sphere, namely: fair opportunities for citizens to participate in public discussions, access to information on matters of public concern that comes from reliable sources, and the equal chance to hear a wide range of views on issues of public concern.[3]

[1] S. Grafanaki, 'Autonomy challenges in the age of big data', *Fordham Intellectual Property, Media and Entertainment Law Journal*, 27 (4), 2016, p. 811.

[2] Dahl argues that a democratic process requires the presence of effective participation, voting equality, enlightened understanding, control of the agenda and inclusion of adults. See R. Dahl, *On Democracy*, (first published by Yale University Press, 1998), Veritas, 2020, pp. 37-38.

[3] J. Cohen and A. Fung, 'Democracy and the Digital Public Sphere', in L. Bernholz, H. Landemore, and R. Reich (eds.), *Digital Technology and Democratic Theory,* The University of Chicago Press, 2021, pp. 29-30.

Although theorists and citizens may disagree about the exact meaning and conditions of democracy, there is widespread concern that various distortions of information caused by digital technologies threaten the functioning of democracy.

As a form of government, democracy relies on citizens' capacity and willingness to participate in the political process. Whereas citizens are assumed to have the general capacity to participate (self-determination), in practice, effective participation may require that citizens have access to a wide range of information and communication opportunities. **Access to information** enables citizens to learn about new issues, form opinions, deliberate and take political action. However, it is not just any type of information that matters. Having plenty of information coming from a single source (e.g. the government or government-controlled media) may not be sufficient. Citizens need access to **'alternative' sources** of information.[4] Lastly, access to diverse information is one of the two aspects of the informational conditions of democracy, the other being the **veracity and reliability of information**. Exposure to a wide variety of false or misleading information would not help citizens to gain 'enlightened understanding' of public issues.

1.2. Key aspects of social media

Social media are online platforms providing 'services that facilitate, organise and amplify the transmission of third-party content, through actions of their registered users'.[5] These platforms provide new opportunities for **expanding and improving democracy**. By removing traditional barriers to creating, transmitting and receiving information, social media can empower citizens to obtain more information, voice opinions, scrutinise government and mobilise for political change.[6]

The rise of online platforms has led to a **paradigmatic shift** in the way information flows in society. Traditional mass media typically broadcast to whole populations and have editors who select and curate information about 'issues, actors, and opinions relevant to society as a whole'.[7] By contrast, social media make a business out of hosting and sharing content created by others. Key differences also exist in terms of accessibility, affordability, speed and reach.[8] Social media platforms are widely accessible and virtually free, and allow users to share content instantly to wide networks of people. They use **sophisticated algorithms** (e.g. newsfeed algorithms and network matching algorithms) that filter content and mediate between producers and recipients of content.

Both the internet and the new digital communication tools have been heralded as great **enablers of democracy**. By circumventing the traditional gatekeepers of information (e.g. mass media, political parties), social media tend to disrupt the political landscape, destabilising governing elites and speeding up the pace of change.[9] Whereas this may have a reinvigorating effect on democracy, it may also enable new actors (e.g. populists) to gain political influence and to use it to undermine democracy. The **global democratic potential** of social media seemed to materialise in the 2000s when successive waves of social protest swept away autocratic regimes around the world. However, the use of social media in this context has been insufficient to bring about real change. As Zeynep Tufekci argued, while social media may enable rapid mobilisation of people, they cannot substitute

[4] R. Dahl, *On Democracy*, p. 85.

[5] J. Bayer, N. Bitiukova, P. Bard, J. Szakács, A. Alemanno, and P. E. Uszkiewicz. Disinformation and propaganda: Impact on the functioning of the rule of law in the EU and its Member States, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, 2019, p.10.

[6] L. Diamond, 'Liberation technology', *Journal of democracy*, 21(3), 2010.

[7] B. Stark and D. Stegmann, 'Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse', *Algorithm Watch*, 2020, p. 11.

[8] C. Wardle and H. Derakhshan, 'Information disorder: Toward an interdisciplinary framework for research and policy making', Council of Europe, 2017, pp. 11-12.

[9] Y. Mounk, *The people versus democracy: The rise of undemocratic liberalism and the threat of illiberal democracy*, Harvard University Press, 2018, p. 149.

for organisation and leadership.[10] Critics have also pointed out that exaggerated enthusiasm about the democratic impact of social media is a form of 'technological solutionism' – the belief that easy technological solutions can fix complex social and political problems[11] – or even an instance of 'techno-narcisism', where outside (Western) commentators celebrate the liberating force of 'their' brilliant technologies while disregarding complex local histories of political struggles.[12]

To better understand how social media affects democracy, five key aspects and issues can be examined:

1. **Surveillance**: social media platforms extract and combine user data to keep users engaged and make profit from selling targeted advertising.
2. **Personalisation**: social media provide personalised content to increase the relevance of information for each user and to bolster engagement.
3. **Disinformation** – social media facilitate the spread of false information either as an unintended consequence or due to certain users' efforts to manipulate the platforms.
4. **Moderation** – social media platforms commonly remove or downgrade content and ban users in order to enforce internal rules and prevent alleged harms.
5. **Micro-targeting** – social media enable targeted advertising that uses granular behavioural data to profile people and to covertly influence their choices.

1.3. Overview of key social media risks to democracy

Table 1 lists 13 specific risks to democracy posed by different aspects of social media. These risks are discussed in more depth and assessed against available evidence in the remainder of the paper. It must be clarified that the analysis does not capture all the risks posed by social media to individuals, institutions, or society (psychological harm, effects on independent press, etc.).

Table 1 – Key social media risks to democracy

Social media aspect	Specific risk	Risk definition
Surveillance	Political surveillance	Social media enable governments to monitor citizens, inhibit their political action and silence dissent
	Loss of privacy and autonomy	Social media surveillance undermines citizens' capacities for political judgement
	Political disengagement	Promotion of viral content and addictive behaviour on social media distracts people away from politics
Personalisation	Narrowed worldviews	Personalised content locks citizens in informational bubbles and affects their capacity to form opinions
	Social and political fragmentation	The segmentation of information and engagement reduces opportunities for political dialogue
Disinformation	Distortion of views and preferences	The spread of false information on social media alters citizens' political views and preferences
	Distortion of electoral outcomes	Disinformation on social media undermines the integrity of elections and affects electoral results
	Automated disinformation	Automated accounts on social media amplify disinformation and exacerbate its effects

[10] Z. Tufekci, Zeynep, *Twitter and tear gas*. Yale University Press, 2017.

[11] E. Morozov, *To save everything, click here: The folly of technological solutionism*, Public Affairs, 2013.

[12] S. Vaidhyanathan, *Antisocial media: How Facebook disconnects us and undermines democracy*, Oxford University Press, 2018, p. 132.

Moderation	Political censorship	Social media moderation undermines freedom of expression and enables control over public opinion
	Algorithmic bias	Automated tools for moderation increase errors, reduce transparency, and automate human bias
	Unaccountable power	Social media platforms take consequential moderation decisions without democratic accountability
Microtargeting	Political manipulation	Microtargeting diminishes citizens' capacities for democratic self-determination
	Distortion of electoral process	Microtargeting distorts electoral process, challenges existing rules, and alters electoral outcomes

2. Surveillance

2.1. Social media and surveillance

Surveillance is the collection and processing of personal data for care or control.[13] Whereas certain surveillance practices are as old as human societies, the more systematic monitoring of populations and individuals is a distinguishable feature of the modern state. Digital surveillance differs from earlier forms of surveillance in several important ways.

Firstly, digital surveillance relies on **more capable technologies** to identify, track, and categorise people. For example, powerful AI algorithms can identify people and predict their behaviour and characteristics by integrating and analysing different streams of data. Highly intrusive technologies are being developed, including face recognition, emotional AI, virtual and augmented reality, and neurotechnology. For example, Facebook is reported to working on developing 'speech decoders' able to determine what people are trying to say by analysing their brain signals.[14] Secondly, governments no longer have monopoly over surveillance. They increasingly **rely on private companies** and 'piggyback'[15] off their surveillance capabilities. Ubiquitous surveillance technologies and multiple surveilling agents make up an 'inscrutable information ecosystem of massive corporate and state surveillance'.[16] Thirdly, digital surveillance can **cover more people** and more aspects of their lives. Digital surveillance targets all people and spaces, and not just specific individuals (e.g. inmates) or places (e.g. public institutions). People's online behaviour and transactions are constantly tracked and analysed. Digital traces are combined to create digital profiles, which are used to predict or infer people's future behaviour, including some of the most intimate aspects of their lives (e.g. sexual orientation, health condition and mood.[17] These digital profiles increasingly determine people's access to information, services, and opportunities.

Social media are part of a new digital industry that relies on sophisticated technology to extract data from online interactions and use it to make a profit by selling advertising – a **business model** that Shoshana Zuboff has termed 'surveillance capitalism'.[18] This business model creates incentives for social media platforms to promote content that is likely to 'trap' peoples' attention[19] and to be shared widely. The more users a platform has and the more active they are on the platform, the more profitable the platform is. As humans tend to be impressed more easily by novel and shocking content, social media algorithms tend to prioritise such content. Whereas the business model of

[13] D. Lyon, *Surveillance society, monitoring everyday life*, Open University Press, 2001.

[14] A. Regalado, 'Facebook is funding brain experiments to create a device that reads your mind', *MIT Technology Review*, 2019.

[15] Grafanaki, 'Autonomy challenges in the age of big data', 2017, p. 815.

[16] Vaidhyanathan, *Antisocial media*, 2018, p. 67.

[17] See, for example, S. Vieira, 'Wake up, algorithms are trawling your phone while you sleep', LSE blog, September 2017.

[18] S. Zuboff, *The age of surveillance capitalism: The fight for a human future at the new frontier of power*, Public Affairs, 2019.

[19] N. Seaver, 'Captivating algorithms: Recommender systems as traps', *Journal of Material Culture*, 24 (4), 2019.

social media is not necessarily geared towards politics, the way in which social media are designed and used may have important (indirect) negative effects on democratic politics.

2.2. Key risks of surveillance

- **political surveillance** – social media enable governments to monitor citizens, inhibit their political action and silence dissent;
- **loss of privacy and autonomy** – surveillance through social media undermines citizens' capacities for political judgement;
- **political disengagement** – the promotion of viral content and addictive engagement on social media distracts people away from politics.

2.2.1. Political surveillance

Social media provide new and more effective ways to monitor people's views, activities, connections, etc. Whereas a lot of surveillance is done for business purposes (e.g. refine advertising strategies), data and insights about persons can be accessed, inferred and sometimes demanded by governments and other parties.

Authoritarian regimes around the globe have been quick to adopt or co-opt digital tools to suppress anti-authoritarian movements and to halt democratisation.[20] Classical political repression techniques, such as street violence and incarceration, are increasingly supplemented by 'smart repression'[21] techniques that make it possible to monitor citizens' activities and silence dissent. Compared to traditional surveillance, digital surveillance has the advantage that it 'requires considerably fewer human actors … entails less physical harassment and comes at a lower cost'.[22]

Evidence of governments **monitoring social media** to keep track of citizens' views and intentions is not hard to come by. According to a 2021 report by Freedom House,[23] about 75 % of internet users worldwide (about 2.8 billion people) live in countries where individuals have been arrested or imprisoned for posting content on political, social, or religious issues. A 2019 report by the Oxford Internet Institute[24] found that computational propaganda was used in 26 countries 'to suppress fundamental human rights, discredit political opponents, and drown out dissenting opinions'. One common tactic is to employ cyber militias and 'troll farms' 'to drown out dissenting voices, accusing them of being "fake news" or "enemies of the people", a sort of censorship through noise'.[25]

The risks of digital surveillance are not only related to non-democratic regimes. In **democratic states** too there are concerns that governments use social media to track people and activities online. Their goals can be legitimate, linked to providing effective public services and security, but also illegitimate, linked to identifying and tracking politically active citizens. For example, technologies to monitor individuals and groups online have been developed to predict political protests following the 2016 US presidential elections.[26]

Digital surveillance may affect political participation even when it is not accompanied by coercive state action. According to the **chilling effect of surveillance**, citizens who expect to be monitored

[20] E. Morozov, *The net delusion: the dark side of Internet freedom*, PublicAffairs, 2012.

[21] L. A. Smithey and L. R. Kurtz, "Smart" repression', in. L. R. Kurtz and L A. Smithey (eds.), *The paradox of repression and nonviolent movements*, Syracuse University Press, pp. 185-214.

[22] S. Feldstein, 'How artificial intelligence is reshaping repression', *Journal of Democracy*, 30(1), p. 42.

[23] Freedom House, 'Freedom on the net 2021: The global drive to control Big Tech', 2021.

[24] S. Bradshaw and P. N. Howard, 'The global disinformation order: 2019 global inventory of organised social media manipulation', Oxford Internet Institute, 2019.

[25] P. Pomerantsev, 'Human rights in the age of disinformation', *Unherd*, 2020.

[26] G. Grill, 'Future protest made risky: Examining social media based civil unrest prediction research and products', *Computer Supported Cooperative Work*, 2021.

may inhibit their political expression and activities online for fear of retribution. On social media, the chilling effects can be amplified by another phenomenon – 'the spiral of silence' – which refers to the process by which individuals self-censor if they believe their opinion belongs to the minority opinion.[27] The chilling effects of surveillance are more difficult to demonstrate, and empirical studies on this topic remain scarce. Several studies found chilling effects in the context of US revelations concerning large-scale government surveillance activities. For example, a study showed how the web traffic to privacy-sensitive Wikipedia articles (e.g. Al Qaeda, fundamentalism) decreased after the publicisation of surveillance practices by the US National Security Agency in June 2013.[28]

2.2.2. Loss of privacy and autonomy

Unwarranted **surveillance interferes with people's privacy** and thus undermine their capacity for democratic self-determination. The link between privacy and democracy is apparent in the basic rule of voting confidentiality during elections. But privacy is important for democracy in a much deeper sense. Democracy relies on citizens' capacity to **think and act autonomously**. However, this capacity does 'not spring full-blown from the womb'.[29] To become autonomous, citizens need a space to experiment and test ideas in a 'zone of relative insulation from outside scrutiny and interference'[30] where they can learn to become autonomous and exercise their autonomy. The right to privacy seeks to provide such a zone of non-interference.

Social media **collect a great amount of user data** usually on the basis that their services are offered for free and that more data are needed to improve these services. While access to data is sometimes provided willingly (e.g. profile information), often users are not aware of the type and amount of data collected. Privacy risks may occur at several levels. Firstly, social media algorithms may collect or share data without the users' explicit consent. Secondly, users' data may be exposed to unauthorised access by third parties. Thirdly, inferences from users' data can be made and used without the users' awareness. Fourthly, the data can be merged and compared with other users' data to create accurate personal profiles even when the data about a particular user is limited.[31]

Individual autonomy may be subverted by deliberate attempts to **undermine people's self-control.** The attention-capture model of social media and other platforms favours addictive behaviour. Algorithms are purposefully designed to trap users into spending more and more time on the networks. For example, research shows that YouTube's recommendation algorithm determines what people watch for more than 70 % of the views on the platform (that is 700 million hours or 1 000 human lifetimes).[32] There are also concerns that, in order to boost engagement, social media platforms actively seek to **exploit human cognitive biases**.[33] For example, they may exploit people's tendency to seek social validation online – for example, by prompting them to like posts that many people liked – and to respond to reward signals, for example, by showering them with constant sound and visual notifications.[34] Social media also benefit from the online disinhibition

[27] M. Büchi, E. Fosch-Villaronga, C. Lutz, A. Tamò-Larrieux, S. Velidi, and S. Viljoen, 'The chilling effects of algorithmic profiling: Mapping the issues', *Computer Law & Security Review*, 36, 2020.

[28] J. W. Penney, 'Chilling effects: Online surveillance and Wikipedia use', *Berkeley Technology Law Journal*, 31(1), 2016.

[29] J. Cohen, *Configuring the networked self: Law, code, and the play of everyday practice*, Yale University Press, 2012.

[30] Grafanaki, 'Autonomy challenges in the age of big data', 2017, p. 811.

[31] S. Milano, M. Taddeo, and L. Floridi, 'Recommender systems and their ethical challenges', *AI & Society*, 35, 2020.

[32] AlgoTransparency, 'Artificial Intelligence (AI) controls what information you are shown on social media: What does it want you to see?', 2021.

[33] G. Sartor, 'New aspects and challenges in consumer protection: Digital services and artificial intelligence', Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020, p. 15.

[34] S. Aral, *The hype machine: How social media disrupts our elections, our economy, and our health – and how we must adapt*, Currency, 2021, p. 106.

effect (stemming from the fact that psychological and social inhibitions tend to drop online because people do not have to take responsibility for their opinions).[35]

2.2.3. Political disengagement

Concerns about the negative effects of communication technologies on democracy are not new. For example, in the 1980s, the American media theorist Neil Postman blamed television for falling levels of citizens' political engagement.[36] Similar objections are raised nowadays with regard to social media. The concern is that, even when they do not spread false information, social media may harm democracy simply by keeping people hooked to their screens. By prioritising content that is likely to impress people, social media may **distract people away from politics**. Moreover, the rise of social media has had a knock-on effect on traditional media pushing them to abandon their civic roles and to shift towards sensationalism, tabloidisation, and 'info-tainment'.[37]

It is true that not so many people are interested in politics or seek political news on social media. For example, in 2020 Facebook reported that political content made only about 6 % of what people in the US see on the network.[38] Moreover, political disengagement and political dissatisfaction among citizens are well-known problems that precede the social media.[39] However, while social media may not be blamed for creating the problem of political disengagement, they may have an important role in **sustaining and aggravating** it. For example, the promotion of shocking stories may trivialise political views, issues and actors, and thus further alienate people from politics.

Political disengagement may not be an intended consequence of social media. It is perhaps unfortunate that people tend to be more interested in 'funny cat videos'[40] than in political debates, but this is not the fault of social media. There is little business incentive to encourage political engagement unless this triggers significant attention (e.g. if users demand that social media take a stand). The main issue here is that social media and their algorithms are **indifferent to democracy**.[41] Given the growing impact of social media on democracy, the question is whether this indifference is sustainable and acceptable.

[35] J. Suler, 'The online disinhibition effect', *Cyberpsychology & Behavior*, 7(3), 2004.

[36] N. Postman, *Amusing ourselves to death: Public discourse in the age of show business*, Viking Penguin, 1985.

[37] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 29.

[38] Facebook, 'What do people actually see on Facebook in the US?', 2020.

[39] There is a well-established literature documenting citizens' dissatisfaction with democracy and political apathy, e.g., R. S. Foa, and Y. Mounk, 'The danger of deconsolidation: the democratic disconnect', *Journal of democracy*, 27(3), 2016.

[40] Historian Yuval Harary complains that we are ready to surrender our personal data 'in exchange for email services and funny cat videos', see Y. N. Harari, *Homo Deus: A brief history of tomorrow*, Random House, 2016, p. 397.

[41] F. Fukuyama, 'Making the Internet safe for democracy', *Journal of Democracy*, 32(2), 2021.

Key points – Risks of surveillance

- social media provide new and more effective ways to monitor people, which can be used by governments to target politically active citizens and silence dissent;
- even in the absence of explicit coercion, citizens who suspect they are the target of online surveillance may inhibit their political expression online for fear of retribution (chilling effects);
- the massive collection of data by social media platforms creates privacy risks to users and may affect their capacity to form and express political opinions;
- the attention capture model of social media, on the one hand, and strategies to exploit human needs and biases to increase engagement, on the other, both undermine individual autonomy;
- social media may contribute to the growing political disengagement of citizens even if they are not solely responsible for this phenomenon;
- the indifference of social media to democracy increasingly clashes with the fact that social media have a growing impact on democracy.

3. Personalisation

3.1. Social media and personalisation

Unlike traditional mass media, which broadcast content to the whole public, social media provide personalised content to each user. The **personalisation of content** can be explicit – when users make explicit choices about their preferred content (e.g. from a pre-defined list), or implicit – when algorithms select relevant content based on users' previous online behaviour.[42] This filtering process may have far-reaching implications because it enables social media algorithms to increasingly re/construct people's subjective reality[43] and to shape the public sphere.

There are clear **benefits of personalisation**. Given the immense volume of information available online, the web would be virtually useless without the ability to sort and rank information. Having access to timely and relevant information can be highly beneficial for citizens and for democracy. However, the empowering potential of personalised information depends greatly on the **quality of information** provided and shared. Without access to relevant and trustworthy information, 'the formation of opinion and thus [citizens'] political decision-making is hindered'.[44]

3.2. Key risks of personalisation

- **narrowed worldviews**: personalised content on social media locks citizens in informational bubbles and affects their capacity to form opinions;
- **social and political fragmentation**: the segmentation of information and engagement on social media reduces opportunities for social and political dialogue.

3.2.1. Narrowed worldviews

The first point of concern about personalisation is that the systematic selection of a certain type of content (and the systematic concealment of other content) presents and reinforces a **narrow and biased worldview** for each user. Social media personalisation may undermine the informational

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

[43] J. Cobbe, and S. Jatinder, 'Regulating recommending motivations, considerations, and principles', *The European Journal of Law and Technology*, 10 (3), 2019.

[44] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 33.

condition of democracy by reducing the variety of information accessible to citizens. For example, an algorithm that recommends users only stories that are similar with what they viewed in the past (e.g. same author, source, ideological orientation) may lock users into information bubbles.

The term '**filter bubbles**' was coined by internet activist Eli Pariser, who contended that personalisation algorithms isolate us into 'a unique universe of information for each of us'.[45] Without access to different opinions, including opinions that contradict their own views, citizens may not be able to develop themselves fully to act autonomously. Personalisation may also **distort citizens' views** about public issues. This is aggravated by the fact that many users are not aware that the information they receive online is pre-selected by algorithms.[46] They may wrongly assume that their views are widely shared by others.

Personalisation algorithms also tend to **reinforce cognitive biases**. For example, algorithms based on pre-selected options would allow people to avoid information that challenges their point of view (selective exposure bias).[47] The systematic encountering of certain type of content (either by choice or by recommendation) could reinforce people's confirmation bias, a phenomenon according to which people are more likely to take note of information that confirms their opinion.

3.2.2. Social and political fragmentation

Apart from creating individual information bubbles, social media algorithms are also blamed for leading to increased convergence of content among people with similar views. Cass Sunstein used the term '**echo chambers**'[48] to describes the effect of group dynamic processes in personalised information environments on individual opinions, in which like-minded individuals constantly reassure themselves of their respective opinions. The concern with online fragmentation goes beyond social media. For example, Google's pageRank algorithm provides different users with different results for the same search terms.[49] The fragmentation of the online world is further amplified by a 'shift from large public groups with open content to private forums, encrypted services, disappearing (or ephemeral) content, and smaller groups'.[50]

The worry is that personalised information increases the fragmentation (or polarisation) and radicalisation of public opinions and disrupts the functioning of the democratic public sphere. The proliferation of deeply personal or parallel online worlds may impede the large-scale political conversations that are necessary for the functioning of democracy, making conflicts more strident and the finding of a compromise more difficult.[51]

The concepts of 'filter bubbles' and 'echo chambers' have been criticised for being vague, unclear, and unsuited for empirical testing.[52] Recent research has shown that fears of filter bubbles and echo chambers are **largely overstated**. For example, a 2019 review of evidence found that the effects of personalisation on the diversity of content available to online users were only minor, if any.[53] A recent (unpublished) paper by researchers at Twitter,[54] which analysed the newsfeeds of about

[45] E. Pariser, *The filter bubble: What the Internet is hiding from you*, Penguin, 2011, p. 9.

[46] G. Marchetti, 'The Role of Algorithms in the Crisis of Democracy', *Athens Journal of Mediterranean Studies*, 6(3), 2020.

[47] Zuiderveen Borgesius et al., 'Should we worry about filter bubbles?', 2016.

[48] C. R. Sunstein, *Echo chambers: Bush v. Gore, impeachment, and beyond*, Princeton University Press, 2001.

.[49] F. Pasquale, *The black box society*, Harvard University Press, 2015, p. 79. For a more recent demonstration, see R. Ochigame and K. Ye, 'Search Atlas: Visualizing Divergent Search Results Across Geopolitical Borders', 2021.

[50] H. Twetman, M. Paramonova, and M. Hanley, 'Social Media Monitoring: A Primer', NATO Strategic Communications Centre of Excellence, 2021, p. 22.

[51] V. Boehme-Neßler, Digitising democracy: On reinventing democracy in the digital era - a legal, political and psychological perspective, Springer, 2020, p. 52.

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

[53] ibid.

[54] F. Huszár, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer & M. Hardt, 'Algorithmic amplification of politics on Twitter', 2021.

two million active users in seven countries, found that the platform's news personalisation algorithms tend to **amplify the mainstream political right** more than the mainstream political left. However, the report found no algorithmic prioritisation of extreme political views (either on the political right of left) over moderate ones.

There are several **mitigating factors** counteracting the potential echo chamber effects of social media. Firstly, an echo chamber effect would require that users are involved in very **homogeneous networks**. This is rarely the case with social media, as they typically involve 'volatile or sporadic contacts with friends from various contexts (e.g. former schoolmates, neighbours, or holiday buddies'.[55] Secondly, although social media are increasingly used, they are (still) **not the only source of information** for most people. So even if social media reinforce selective exposure and confirmation bias, the result 'might not be as dramatic as often suggested. Even if people self-select consonant content, they may still be confronted with conflicting content.[56] Thirdly, the political effects of social media algorithms rely on the premise that social media are a major **source of political information**. However, research show that many social media users rarely encounter political information on the platforms, and that they usually do not choose their network of friends according to political views. For example, a 2021 study[57] analysing Twitter messages shared by 1.4 million US users found that in-group messages were shared 14 times more frequently than outgroup messages. However only about 40 % of these users were interested in politics.

Whereas the negative political effects of social media personalisation seem less severe and widespread, the **risk of societal fragmentation** and political polarisation remains. There is evidence that social networks push people to cluster more than they would usually do in the offline world (the homophily effect), for example, through friend-suggestion algorithms.[58] There is also evidence of 'echo chamber' effects among hyperpartisans at the fringes of the political spectrum. As a 2021 report[59] by researchers from New York University concludes, although social media may not be the main cause of polarisation (in the US), the 'use of these platforms **intensifies divisiveness**'. There is also research showing that the political polarisation effect of social media **depends on the political context**. For example, whereas polarisation patterns have been identified in bi-partisan political systems, such as in the UK or the US, they seem to be less obvious in multiparty systems.[60]

It must be noted that political polarisation is a **contested issue**. This has to do with the fact that democracy itself is a contested concept. As Dahl pointed out, democracy 'has meant different things to different people at different times and places' and, despite being discussed and practised for 25 centuries, we still have no definitive agreement on 'the most fundamental questions about democracy'.[61] Policy discussions about the influence of social media on democracy tend to treat democracy as a **self-evident ideal**, although the nature and form of democracy can be conceptualised in a range of different ways.[62] For example, theorists as well as citizens may weigh differently the importance of the public sphere and the risks of polarisation. On the one hand, political polarisation is regarded as problematic, because it makes political compromise more

[55] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 27.

[56] Zuiderveen Borgesius et al., 'Should we worry about filter bubbles?', 2016, p. 6.

[57] M. Wojcieszak, A. Casas, X. Yu, J. Nagler, and J. A. Tucker, J.A., 'Echo chambers revisited: The (overwhelming) sharing of in-group politicians, pundits and media on Twitter', 2021.

[58] Aral, *The hype machine*., 2021, p. 71.

[59] P. M. Barrett, J. Hendrix, and J. G. Sims, 'Fueling the fire: How social media intensifies U.S. political polarization – and what can be done about it', New York University, 2021.

[60] Zuiderveen Borgesius et al., 'Should we worry about filter bubbles?', 2016, p. 8.

[61] Dahl, *On Democracy*, p. 11.

[62] B. Barrett, K. Dommett, and D. Kreiss, 'The capricious relationship between technology and democracy: Analyzing public policy discussions in the UK and US', *Policy & Internet*, 2021.

difficult and in extreme cases impossible to achieve.[63] On the other hand, there are theorists challenging consensualist approaches to democracy and claiming that conflict and polarisation are acceptable and even desirable for a democracy.[64]

Key points – Risks of personalisation

- whereas content personalisation can help citizens deal with the problem of information overload, it can also narrow their worldviews by limiting the range of information available to them;
- the patterns of information-sharing and interactions on social media can contribute to the reinforcement of group boundaries that would preclude wider political dialogue;
- despite widespread concerns, existing empirical evidence suggests that the personalisation and filtering effects of social media are less severe and pervasive that initially feared;
- the negative effects of personalisation are more prevalent in certain political contexts and tend to affect more certain groups, such as hyperpartisans;
- whereas the negative political effects of personalisation seem less severe and widespread, the risk of societal fragmentation and polarisation remains;
- the assessment of the effects of social media depends also on political (ideological) assumptions about the nature and conditions of democratic politics.

4. Disinformation

4.1. Social media and disinformation

Disinformation is a type of information disorder where false information is intentionally spread to cause harm, deceive, or obtain economic gain.[65] Social media offer great opportunities to **spread false information to a great number of people**. There are various methods and tools used to spread disinformation through social media, including by artificially amplifying false stories (through hashtags and links), using automated accounts (bots) to generate posts or engage with content, masking the real source or sponsor of a message (astroturfing), impersonating authoritative media, people or governments, and using digitally altered or fabricated videos or audio (deep fakes).[66]

Promoting sensationalist news, conspiracy theories, and unrealistic political messages is not a new phenomenon. Evidence of spreading disingenuous news is as old as the cuneiform tablets of Hammurabi.[67] But social media (and other online platforms) offer a much more **powerful, accessible and cheap** tool for promoting false information than previous means. Several characteristics of the online environment make social media platforms a fertile ground for disinformation: wide reach (potentially global), interactivity (users are active producers of information as opposed to passive consumers), and immediacy (almost instant communication).[68]

[63] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 15.

[64] J. Cowls, 'Deciding how to decide: Six key questions for reducing AI's democratic deficit', *Medium*, 2019.

[65] European Commission, 'A multi-dimensional approach to disinformation', High level Group on fake news and online disinformation, 2018.

[66] C. Colomina, H. Sánchez Margalef, and R. Youngs, 'The impact of disinformation on democratic processes and human rights in the world', Policy Department for External Relations, European Parliament, 2021, p. 7.

[67] C. Marsden, and T. Meyer, 'Regulating disinformation with artificial intelligence', EPRS, European Parliament, 2019. For a brief history of 'fake news', see D. Uberti, 'The real history of fake news', *Columbia Journalism Review*, 2016.

[68] Twetman et al., 'Social media monitoring', 2021, p. 10.

The **term 'fake news'** has been widely used to describe various kinds of information distortion. However, critics have pointed out that the term is ambiguous and inadequate to describe information distortion.[69] Moreover, it has also been 'appropriated by politicians around the world to describe news organizations whose coverage they find disagreeable'.[70] A more useful approach is to distinguish between three types of information disorder: misinformation – when false information is shared but no harm is meant, disinformation – when false information is knowingly shared to cause harm, and malinformation – when genuine information is shared to cause harm.[71]

Whereas the political effects of personalisation may be a side effect of social media, disinformation is a **deliberate action**. A clarification needs to be made here: disinformation is by no means always online – there is still plenty of offline disinformation, and it is not always (strictly) political (as in the case of issues relating to vaccines, climate, etc.). Moreover, whereas some disinformation campaigns are motivated by political ends, a lot of disinformation is driven purely by economic rationale. For example, a group of teenagers in a small town in North Macedonia reportedly set up hundreds of websites to promote false news during the 2016 US presidential elections, which brought them over US$20 million a year in ad revenue.[72]

4.2. Key risks of disinformation

- **distortion of views and preferences** – the spread of false information on social media alters citizens' political views and preferences;
- **distortion of electoral outcomes** – disinformation on social media undermines the integrity of elections and affects electoral results;
- **automated disinformation** - automated accounts on social media amplify disinformation and exacerbate its effects.

4.2.1. Distortion of views and preferences

The spread of false information on social media undermines citizens' capacity to form and express political views and inhibits the free formation of public opinion.[73] When exposed to disinformation, citizens may uncritically form or change their political views and preferences based on false information or false perceptions of other peoples' opinions.

There is growing empirical research showing significant **exposure to political disinformation** on social media. For example, Russian 'trolls' spreading false information during the 2016 US presidential elections are believed to have reached up to 126 million Facebook users, over 20 million Instagram users, and 1.4 million Twitter users.[74] A 2020 report by the Oxford Internet Institute[75] found 'evidence of 81 countries using social media to spread computational propaganda and disinformation about politics.' Although research on Europe is less voluminous, evidence of exposure to online disinformation during elections was unveiled in France (2017), Germany (2017), and Italy (2018).[76]

Online disinformation is not caused only by **foreign governments** interfering in other countries. In fact, **threats from domestic actors** attempting to undermine democracy from within are becoming

[69] P. Müller, and N. Denner, 'What can be done to counter fake news', Friedrich Naumann Foundation for Freedom, 2019.

[70] Wardle and Derakhshan, 'Information disorder', 2017, p. 5.

[71] ibid.

[72] S. Subramanian, 'The Macedonian teens who mastered fake news', *Wired*, 2017.

[73] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 32.

[74] S. McKay, and C. Tenove, 'Disinformation as a threat to deliberative democracy'. *Political Research Quarterly*, 74(3), 2021; D. Ingram, 'Facebook says 126 million Americans may have seen Russia-linked political posts', *Reuters*, 2017.

[75] Bradshaw and Howard, 'The global disinformation order', 2019.

[76] M. Cantarella, N. Fraccaroli, and R. Volpe, 'Does fake news affect voting behaviour?', CEIS Working Paper 493, University of Rome Tor Vergata, 2020.

more visible, and the lines separating foreign from domestic interference are becoming increasingly blurred.[77] NATO's 2019 report on social media and disinformation[78] found that it was still cheap and easy for foreign governments and anti-democratic groups to interfere with voters' choices and run manipulative social media campaigns. The 2020 edition of the report concluded that 'despite significant improvements by some, none of the five platforms [Facebook, Instagram, Twitter, YouTube, and TikTok] is doing enough to prevent the manipulation of their services'.[79]

The political impact of disinformation is **difficult to quantify**. One major challenge is that disinformation usually targets people who are likely to be susceptible to it (selection effect).[80] As disinformation moves from social media to closed communication environments, such as encrypted messaging apps, it becomes even more difficult to assess its impact.[81]

Existing evidence show that the reach and impact of disinformation are **likely to be overestimated**.[82] According to a 2021 paper, 'the few studies that have empirically tested the reach of disinformation consistently find this reach to be severely limited'.[83] Moreover, the exposure to and engagement with false content online seem to **vary greatly across groups** and individuals. For example, a study on the flow of disinformation on Facebook during the 2016 US presidential election found that 'on average, users over 65 shared nearly seven times as many articles from fake news domains as the youngest age group'.[84] The study concluded that the effects of false news articles on citizens' political attitudes are likely dampened because only a specific small group of citizens is exposed to false information.

Other studies found that 'disinformation taps into **pre-existing attitudes** that are confirmed and (moderately) strengthened'.[85] For example, a study[86] on the effects of false information on populist voting in the 2018 Italian election in the region of Trentino Alto-Adige in South Tyrol found that 'voters self-select into misinformation bubbles and consume fake news because of their prior preference for populist platforms, and not the other way around.' Another study,[87] measuring the impact of disinformation on the political attitudes and behaviours of US Twitter users in late 2017, found no evidence of Russian trolls affecting Americans' political attitudes, beyond those who 'were already highly polarized.'

While somewhat reassuring, the evidence of the limited impact of disinformation on citizens' political views does not dispel concerns about broader challenges posed by disinformation on democratic process. For example, it is argued that disinformation **undermines trust in democratic institutions** by creating 'a trail of doubt as to whether democratic institutions actually work well in

---

[77] N. Bentzen, Trump's disinformation 'magaphone': Consequences, first lessons and outlook, briefing, European Parliament, EPRS, 2019.

[78] R. Fredheim, and S. Bay, 'How social media companies are failing to combat inauthentic behaviour online', NATO Strategic Communications Centre of Excellence, 2019, p. 19.

[79] R. Fredheim, S. Bay, A. Dek, and I. Dek, 'Social Media Manipulation Report 2020,' NATO Strategic Communications Centre of Excellence, 2020.

[80] Aral, *The hype machine*., 2021, p. 141.

[81] P. Rossini, J. Stromer-Galley, E. A. Baptista, and V. Veiga de Oliveira, 'Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections', *New Media & Society*, 2020.

[82] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 4.

[83] A. Jungherr, R. Schroeder, 'Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy', *Social Media + Society*, 2021.

[84] A. Guess, J. Nagler, and J. Tucker, 'Less than you think: Prevalence and predictors of fake news dissemination on Facebook', *Science advances* 5(1), 2019.

[85] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 37.

[86] M. Cantarella *et al*., 'Does fake news affect voting behaviour?', 2020, p. 27.

[87] C. A. Bail, B. Guay, E. Maloney, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, and A. Volfovsky, 'Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017', *Proceedings of the national academy of* sciences ,117(1), 2020.

reflecting citizens' choices.[88] Moreover, a strong **public perception** of widespread disinformation can affect trust even if the actual reach and impact of disinformation are low.

According to the latest results of the Flash **Eurobarometer on fake news** and disinformation online,[89] 68 % of Europeans declared that they had come across news or information that they believe misrepresented reality or was false (37 % believed they encountered this every day or almost every day). Some 69 % of respondents were confident that they could identify misleading or false news; 28 % were not. The overwhelming majority (85 %) considered the existence of news or information that misrepresent reality or is even false a problem. A 2021 EU citizens' survey on democracy[90] also showed that a majority of internet-using Europeans (51 %) believed they had been exposed to or personally witnessed disinformation on the internet; 45 % of internet users considered that they had been exposed online to content created to divide society on a specific issue. According to Reuter's 2020 Digital News report,[91] which surveyed people in 40 countries (including 21 EU countries), people see social media (in particular Facebook) as the biggest source of concern about misinformation (40 %), well ahead of news sites (20 %), messaging apps (14 %), and search engines (10 %). These perceptions, however, vary across countries, including in Europe.

Whereas an acute perception of disinformation can indicate that citizens are aware of the phenomenon and thus less likely to be deceived by it, this perception may also be a sign of generalised mistrust in online information, including truthful information from legitimate sources. One key point here is the need for more and better research on the impact of disinformation and, importantly, for **accurate and responsible reporting** of such findings.

4.2.2. Distortion of electoral outcomes

Another concern is that widespread disinformation can compromise the integrity of elections and **distort electoral outcomes**. There are different ways in which disinformation can affect elections. First, citizens may be persuaded to change their political preferences based on false information or false perceptions of other peoples' opinions. Second, disinformation can be used to mobilise or demobilise people to cast their vote. There is research showing that ideologically consistent false information 'can motivate voters to turn out even if it doesn't change their vote choices' and that 'targeted efforts to increase or diminish voter turnout could be substantial enough to change overall election results'.[92]

The electoral effect of disinformation may also depend on the **political system and the type of election or voting**. In bi-partisan political systems there is a greater chance for a small minority of voters to have a decisive contribution on the results of an election. So even if disinformation is not as widely spread and even if it does not influence a great number of people, it can still change electoral outcomes by persuading a small minority of voters.

Early research conducted by Facebook showed that social media messages encouraging people to vote could have a significant **impact on political mobilisation**. For example, during the 2010 US mid-term congressional elections, Facebook posted messages such as 'I voted' or 'Find your polling place' on the newsfeed of 61 million US users. It then estimated that a single such message triggered the casting of 340 000 additional votes.[93] In a follow-up experiment during the 2012 US presidential elections, Facebook messages targeting 15 million US user triggered 270 000 additional votes each. This may be a small change in a country as big as the US, but enough to make a difference. For

[88] Colomina et al., 'The impact of disinformation on democratic processes', 2021, p. 15.

[89] Flash Eurobarometer 464: Fake news and disinformation online, updated 4 May 2021.

[90] Special Eurobarometer 507: Democracy in the EU, report, 2021.

[91] N. Newman, F., Richard, A. Schulz, S. Andi and R. Kleis Nielsen, 'Digital news report 2020', Reuters Institute and University of Oxford, 2020, pp. 18-19.

[92] Aral, *The hype machine*, 2021, p. 35 and 37.

[93] ibid., pp. 156-7.

example, in the 2000 presidential election, George W. Bush defeated Al Gore with a margin of just 537 votes in Florida. Facebook has displayed voter mobilisation messages to European users too, including in relation to the 2014 Scottish referendum, the 2015 Irish referendum, and the 2016 UK Brexit referendum. Whereas political mobilisation is generally a positive thing, these findings show the significant **power of social media** to change citizens' political behaviour. When used for less noble purposes (for instance, for disinformation and manipulation), this power can be destructive for democracy.

The electoral impact of disinformation has been discussed extensively in the context of the 2016 US presidential election and the UK Brexit referendum. In Europe, disinformation has been held responsible for the recent electoral successes of **populist parties**.[94] A 2020 study[95] focused on the 2017 German parliamentary election found that voters who were exposed to disinformation were more likely to be alienated from the mainstream political parties and pushed towards populist parties. Interestingly, the study also revealed that 'the less one trusts in news media and politics, the more one believes in online disinformation'.

Another risk is that non-democratic forces, including **foreign governments**, may use disinformation campaigns to exert undue influence on democratic processes and elections. This 'weaponization of disinformation'[96] is a key challenge to democracy. For example, a 2019 report[97] for the US Senate found that the Russian government supported the interference in the 2016 US election as 'part of a broader, sophisticated, and ongoing information warfare campaign designed to sow discord in American politics and society'. It must be noted that this and similar studies reveal the significant supply of disinformation but do not say much about the actual impact of disinformation on voting.

Even if disinformation does not substantially distort citizens' opinions or alter electoral outcomes, the fact that false political information shared through social media reaches such a large audience may have a stronger impact on democracy. Disinformation may **undermine trust** in democratic institutions, accelerate political disengagement, and polarise society by sharpening 'existing sociocultural divisions using nationalistic, ethnic, racial and religious tensions'.[98]

4.2.3. Automated disinformation

Algorithms and automation play a key role in spreading disinformation though social media. On the one hand, **filtering algorithms** used for personalisation and for boosting engagement may increase the visibility of content that contains false or misleading information. On the other hand, users and **disinformation 'entrepreneurs'** may deliberately seek to trick algorithms to spread and amplify the visibility of false information.[99]

Automated accounts or bots can achieve scalability of disinformation by massively spreading false information.[100] They can also enhance the credibility of a message by artificially inflating the perceived popularity of a story or account by generating fake reactions, comments or followers. For

[94] Cantarella *et al*., 'Does fake news affect voting behaviour?', 2020.

[95] F. Zimmermann, and M. Kohring, 'Mistrust, disinforming news, and vote choice: A panel survey on the origins and consequences of believing disinformation in the 2017 German parliamentary election', *Political Communication*, 37(2), 2020

[96] European Political Strategy Centre, 10 trends shaping democracy in a volatile world, 2019.

[97] US Senate, 'Report on Russian active measures, campaigns and interference in the 2016 U.S. election, volume 2: Russia's use of social media, 2020.

[98] Wardle and Derakhshan, 'Information disorder', 2017, p. 4.

[99] N. Maréchal, 'Automation, algorithms, and politics| when bots tweet: Toward a normative framework for bots on social networking sites (feature)', *International Journal of Communication*, 10, 2016.

[100] M. Brkan, 'Artificial intelligence and democracy: The impact of disinformation, social bots and political targeting', *Delphi*, 2, 2019.

example, sockpuppets or trolls pretend to be ordinary human users to influence real social network users by creating the impression that certain views have wide public support.[101]

More sophisticated **artificial intelligence** (AI) technologies making it possible to manipulate images, video (deep fakes) and text in order to create highly realistic depictions of real people doing or saying things could significantly increase the credibility of false information.[102] These could be used to 'discredit leaders and institutions, incite violence and tilt cities towards civil unrest, exacerbate existing divisions in society, or influence the outcome of elections'.[103] For example, a recent study[104] showed that the OpenAI's language model for automated content creation, GPT-3, can already be an effective tool to create messages for large-scale disinformation campaigns, in particular when paired with a skilled operator and editor.

Evidence of bots spreading false information has been revealed in relation to the Brexit referendum campaign, the debates on the UN migration pact in Germany, the 2017 German elections, the 2018 Swedish elections, the 2017 French presidential elections, and the 2017 Catalan independence referendum.[105] However, the **prevalence or influence** of social bots on social media has been questioned. A recent paper[106] argued that research showing a high number of social bots acting autonomously on social media to spread false information has important methodological flaws. Even if social bots play a role in disinformation, their influence seem to be concentrated mainly in the early stages of disinformation spreading.[107] Online disinformation requires an **interaction between users and algorithms**. Research on the dynamics of online disinformation show that bots are typically the initial high spreaders of false information, which is then propagated knowingly or unknowingly by humans.[108] One fact that explains the 'success' of disinformation online is that human attention is more easily attracted by novel information (the 'novelty hypothesis').[109] In a 2018 study of rumour cascades on Twitter, researchers found that **false news spreads much faster and broader** than true stories and suggested that this was caused by the degree of novelty involved and the strong emotional reactions of recipients.[110]

[101] A. Alaphilippe, A. Gizikis, C. Hanot, and K. Bontcheva, 'Automated tackling of disinformation. Major challenges ahead', EPRS, European Parliament, 2019.

[102] N. Giansiracusa, *How algorithms create and prevent fake news: Exploring the impacts of social media, deepfakes, GPT-3, and more*, Apress, 2021; S. Kreps, and M. McCain, 'Not your father's bots: AI is making fake news look real', *Foreign Affairs*, 2019.

[103] R. Chesney, and D. Citron, 'Deepfakes and the new disinformation war: The coming age of post-truth geopolitics', *Foreign Affairs*, 2019.

[104] B. Buchanan, A. Lohn, M. Musser, and K. Sedova, 'Truth, lies, and automation: How language models could change disinformation', Georgetown University, 2021.

[105] Brkan, 'Artificial intelligence and democracy', 2019, p. 3.

[106] Gallwitz and Kreil, 'The rise and fall of 'social bot' research', 2021.

[107] Alaphilippe et al., 'Automated tackling of disinformation', 2019, p. 34.

[108] Aral, *The hype machine*, 2021, p. 48.

[109] ibid., p. 49.

[110] V. Soroush, D. Roy, and S. Aral, 'The spread of true and false news online', *Science*, 359, 2018.

Risks of disinformation – Key points

- whereas disinformation is sometimes motivated by political ends, a lot of it is driven by entrepreneurs promoting highly engaging content to make profits from selling ads
- the spread of false information on social media can undermines citizens' capacity to form and express political views;
- despite growing evidence of significant exposure to political disinformation online, the actual impact of disinformation on citizens' views and preferences is difficult to assess;
- although the reach and impact of disinformation seem to have been overestimated, there is evidence of negative effects in particular contexts and on specific groups;
- disinformation can be used to persuade or confuse voters, and to mobilise or demobilise citizens to cast a vote, which may, in certain conditions, be a determinant of election outcomes;
- disinformation is not only caused by foreign actors (e.g., governments) seeking interference in other countries, but also by domestic actors;
- widespread disinformation and acute public perception thereof (amplified by lack or research and inadequate reporting) may undermine trust in (all) online information and democratic institutions;
- automated accounts and algorithms contribute to the spread of disinformation on social media, but interaction between users and algorithms is a must if disinformation is to be effective;
- algorithms used to spread false information exploit human biases and predispositions, such as human confirmation bias, inclination to believe repeated stories, and attraction to novel content.

5. Moderation

5.1. Moderation by social media platforms

Moderation refers to 'the detection of, assessment of, and interventions taken on content or behaviour deemed unacceptable by platforms or other information intermediaries, including the rules they impose, the human labour and technologies required, and the institutional mechanisms of adjudication, enforcement, and appeal that support it'.[111] Over the years, crude moderation mechanisms such as blocking content and banning accounts have developed into a more complex set of tools that include: quarantining topics, removing posts from search, barring recommendations, down-ranking posts in priority, verifying content and labelling.

Moderation **practices are common** on social media platforms. Some instances of moderation are highly visible, for example when these involve labelling content or suspending accounts of political leaders (e.g. Donald Trump's suspension from major platforms in January 2021[112]). However, a lot of content moderation by platforms remains invisible.[113] For example, it is reported that, in the first half of 2020 alone, Twitter suspended roughly 925 000 accounts for rules violations. Although moderation on social media is **not primarily driven by political concerns**, moderation practices may significantly affect democracy, for example, by limiting citizens' right to information and to freedom of expression.

[111] T. Gillespie, P. Aufderheide, E. Carmi, Y. Gerrard, R. Gorwa, A. Matamoros-Fernández, S. T. Roberts, A. Sinnreich, and S. Myers West, 'Expanding the debate about content moderation: scholarly research agendas for the coming policy debates', *Internet Policy Review*, 9(4), 2020, p. 2.

[112] N. Bentzen, Trump's disinformation 'magaphone', 2019.

[113] M. Luca, 'Social media bans are really, actually, shockingly common', *Wired*, 2021.

5.2. Key risks of moderation

- **political censorship** – moderation by social media platforms undermines freedom of expression and enables platforms and governments to control public opinion;
- **algorithmic bias** – automated moderation tools increase errors, reduce transparency and automate human bias;
- **unaccountable power** – social media platforms take consequential moderation decisions without democratic accountability.

5.2.1. Political censorship

The most prominent social media platforms have recently become more aware of the adverse impact of the spread of false and harmful information through their networks and have started using moderation to tackle these issues. Even though such political moderation is often pursued in the name of protecting democracy, the policing of content and users on social media poses important challenges to democracy. Such measures may even 'pose a greater harm to democracy than disinformation itself'.[114]

Freedom of expression is both an **individual rights and a core value of democracy**. Social media and other online platforms provide important and sometimes unique opportunities for freedom of expression. As a 2015 judgement of the European Court of Human Rights states, social media (in this case, YouTube) provide a 'unique' and 'undoubtedly an important means of exercising the freedom to receive and impart information and ideas', allowing the expression of some political content that may otherwise be 'ignored by the traditional media'.[115] Although the right to disseminate and access information is not limited to true information, there may be instances in which the right to freedom of expression **may be limited**. For example, according to Article 17 of the European Convention on Human Rights (ECHR), the ECHR does not protect any activity aimed at the destruction of any of the rights and freedoms contained in the ECHR, such as speech that endangers free operation of democratic institutions or attempts to destroy the stability and effectiveness of a democratic system.[116] However, any limitations of the right to freedom of expression need to be justified.

This risk of political censorship increases when moderation is performed by private companies **acting on the instruction of government**. Platform oversight could be co-opted by (authoritarian) governments to censor their citizens, criminalise journalism and undermine fundamental rights.[117] For example, in 2021, Freedom House reported[118] 'a record-breaking crackdown on freedom of expression online' globally. The report documented people being arrested or convicted for their online speech in 56 countries, and governments in 21 countries having blocked access to social media platforms, most often during political protests and elections.

Some moderation practices are more problematic than others. Whereas removing inauthentic accounts may be acceptable for the purpose of increasing transparency and avoiding manipulation, **removing content or blocking users** from posting content raises serious concerns about censorship and interference with freedom of expression.[119] The removal of certain content from the

[114] Bayer et al., 'Disinformation and propaganda – impact on the functioning of the rule of law', 2019, p. 10.

[115] European Court of Human Rights, 'Case *Cengiz and Others v. Turkey*', Judgement of 1 December 2015, para 52; See also M. van Riel, 'Cengiz and Others v. Turkey: a tentative victory for freedom of expression online', *Strasbourg Observers*, 2016.

[116] A. Kuczerawy, 'Does Twitter trump Trump?: A European perspective', *VerfBlog*, 2021.

[117] J. Hand, "Fake news" laws, privacy & free speech on trial: Government overreach in the infodemic?', *First Draft*, 2020.

[118] Freedom House, 'Freedom on the net 2021'.

[119] Amnesty International, 'Surveillance giants: How the business model of Google and Facebook threatens human rights', 2019.

public sphere 'may silence some speakers (e.g. social activists, political opponents) and may deprive the public of access to legitimate speech'.[120]

Social media have the obligation to **remove illegal content** under different national and EU laws, but there is increasing pressure on social media platforms to take action against content that is considered harmful although not explicitly illegal. This may create additional risks. As the authors of a 2019 study[121] argue, 'restrictions to freedom of expression must be provided by law, legitimate and proven necessary, and as the least restrictive means to pursue the aim'.

As an alternative to removing content, social media can tweak their algorithms to **reduce the power of problematic content** (e.g. update news feed algorithms to de-emphasise disinformation). For example, Facebook is currently testing reducing visibility of all political news seen by users in several countries, including Ireland, Spain, and Sweden.[122] While this is done in response to users' feedback, deprioritising political content across the board may also reduce citizens' opportunities to access political content, particularly in times when this is most needed, such as during elections. While deprioritisation measures could help filter out malicious content,[123] they can also be used by social media platforms 'to manipulate public opinion.'[124]

Another way to deal with disinformation on social media is to **check and label content** using source transparency indicators. This has the advantage that it involves users in the process, while also enabling them to understand how their social media feeds are built and to edit their feeds.[125] However, efficient correction and labelling of false information is difficult and costly. The **effectiveness** of this approach has also been questioned. According to some research, trying to correct people's false beliefs by labelling or discrediting content could back-fire and push people to reinforce their opinions and beliefs (backlash effect) or to doubt the accuracy of all headlines, including true ones.[126] People may also consider accurate all untagged headlines, including false ones ('implied truth effect').[127] The 'backlash' effect has nevertheless been challenged by research.[128]

Deleting and labelling certain content may also be **counter-productive**, as populists and peddlers of conspiration theories can point to these efforts as proof that elites are working hard to supress popular views.[129] Moreover, removing content and blocking accounts on dominant platforms may also push people to move to smaller and potentially more radical platforms, thus possibly increasing political fragmentation and polarisation. For example, there is evidence that many conservative-oriented users in the US moved to smaller platforms such as Parler, Gab, MeWe and Locals, when big platform started to enforce stronger moderation.[130]

---

[120] N. Elkin-Koren, 'Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence', *Big Data & Society*, 7(2), 2020, p. 4.

[121] Marsden and Meyer, 'Regulating disinformation with artificial intelligence', 2019.

[122] A. Gupta, 'Reducing political content in news feed', *Facebook news*, 2021.

[123] N. Maréchal and E. R. Biddle, 'It's not just the content, it's the business model: Democracy's online speech challenge', *New America*, 2020.

[124] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, and A. Dafoe, 'The malicious use of artificial intelligence: Forecasting, prevention, and mitigation', 2018.

[125] Alaphilippe et al., 'Automated tackling of disinformation', 2019, p. 34.

[126] K. Clayton, S. Blair, J.A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, and M. Sandhu, 'Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media', *Political Behavior*, 42(4), 2019.

[127] G. Pennycook, A, Bear, E. T. Collins, and D. G. Rand, 'The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings', *Management Science*, 66(11), 2020.

[128] A. Guess, and A. Coppock, 'Does counter-attitudinal information cause backlash? Results from three large survey experiments', *British Journal of Political Science*, 50(4), 2020.

[129] Müller and Denner, 'What can be done to counter Fake News', 2019.

[130] Bentzen, Trump's disinformation 'magaphone', 2019.

5.2.2. Algorithmic bias

Content moderation is not only hard and risky but it also tends to yield limited results. Manual (human) fact-checking is a complex and time-consuming process and it can therefore only cover a small proportion of social media content. Automated tools and algorithms[131] promise to address the **problems of scale and resources** linked to moderation. Automation can also alleviate the **psychological cost** of human moderators. Human moderators often work in highly stressful conditions and on tight schedules, and can struggle to cope with traumatic images and videos.[132] Lastly, certain types of content can only be detected with automated tools, as is the case of AI-generated deep-fakes.[133]

Automation poses specific risks of **errors and bias**. Moderation algorithms make mistakes, which can be costly. For example, false negatives (when disinformation is wrongly labelled as true or bot accounts are wrongly identified as human) and false positives (when correct information is wrongly labelled as disinformation or genuine users are wrongly identified as bots) may result in wrongly suspended accounts, failure to sufficiently prevent the spread of hate and online abuse, and increased exposure to junk news.[134] Revelations from the 2021 'Facebook papers' show that the social media platforms are well aware of the ineffectiveness of automated systems.[135]

Automated tools may discriminate against people who speak/write in a certain way because the technology has **limited capacity to understand context**. For example, research in the US context has shown that algorithms trained to identify hate speech for removal were more likely to flag social media content created by certain groups or minorities, such as African Americans using slang.[136] Automated tools have trouble 'parsing multiple, complex, and possibly conflicting meanings emerging from text'.[137] While content moderation algorithms can identify a distinct symbol, such as a swastika, they have difficulties assessing content that is 'violent, hateful, or misleading and yet has some public interest value', a job that is best suited for human judgement.[138]

Another problem is that it is often difficult to assess the accuracy of algorithms, as many of these are proprietary algorithms that are not open to public scrutiny (**black boxes**). This means that we will not be able to easily assess whether and when algorithms replicate and automate human biases.

Lastly, mandating social media platforms to take on greater responsibilities for tackling harmful content may push them to over-zealously remove what seems to be problematic content for fear of penalties, thus leading to overcensorhip.[139] The more heavily social media **rely on automated tools**, the greater the risk of overcensorship might be. Increased ex-ante filtering and monitoring of content 'may lead to an undesirable compression of users' fundamental freedoms and rights, such as freedom of expression'.[140]

---

[131] For overviews and discussions of automated tools to fighting online misinformation, see Alaphilippe et al., 'Automated tackling of disinformation', 2019; H. Twetman et al., 'Social media monitoring', 2021; and RAND, 'Fighting disinformation online: A database of web tools', 2019.

[132] S. Roberts, *Behind the screen: Content moderation in the shadows of social media*, Yale University Press, 2019. See also, C. Newton, 'The trauma floor: The secret lives of Facebook moderators in America', *The Verge*, 2019.

[133] Brkan, 'Artificial intelligence and democracy', 2019.

[134] Alaphilippe et al., 'Automated tackling of disinformation', 2019.

[135] D. Seetharaman, J. Horwitz and J. Scheck, 'Facebook says AI will clean up the platform. Its own engineers have doubts', *Wall Street Journal*, 17 October 2021.

[136] M. Sap, D. Card, S. Gabriel, Y Choi, and N. A. Smith, 'The risk of racial bias in hate speech detection', Proceedings of the 57th annual meeting of the association for computational linguistics, 2019.

[137] Marsden and Meyer, 'Regulating disinformation with artificial intelligence', 2019.

[138] Maréchal and Biddle, 'It's not just the content, it's the business model', 2020.

[139] K. Kertysova, 'Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered', *Security and Human Rights*, 29 (1-4), 2018, p. 73.

[140] A. Bertolini, H. Episcopo, and N.A.Cherciu, Liability of online platforms, EPRS, European Parliament, 2021, p. 55.

5.2.3. Unaccountable power

Following the decisions of major social media platforms in early 2021 to ban Trump's accounts, policymakers and commentators around the world have reiterated their concern about the power of private platforms to censor political speech. Whereas such bans may not have been illegal,[141] they raise crucial questions about the **legitimacy** of these platforms' power to regulate public speech and 'protect' democracy.

Social media platforms typically constrain the type of content users can post on their platforms through their terms of use and community standards. However, these rules are often **unclear, arbitrary**[142] **and inconsistently applied**. For example, Twitter was accused of following a double standard by banning Trump from the platform while allowing autocrats in other parts of the world to use it to harass opponents.[143]

Despite recent commitments and measures, there are reports that **inconsistencies prevail** in how the largest platforms respond to disinformation. The 2020 NATO report on social media and disinformation points outs that some platforms are better than others, with newer and smaller platforms being 'nearly defenceless against platform manipulation'.[144] Inconsistencies may be caused by the fact that social media's **business priorities** (e.g. increasing user engagement) are often in conflict with their public commitments (e.g. tackling disinformation).[145] The risk is that platforms may prioritise addressing certain issues, which may not necessarily be the most important ones from a disinformation containment perspective.[146]

Apart from issues of accuracy and transparency, the thorny question about moderation is **who decides** what problematic content is. To answer this question, Facebook has appointed an independent Oversight Board, composed of legal and human rights experts, which adjudicates on what should be allowed on or removed from the platform (sensitive cases). While this presumably takes the matter out of the hands of the platform, it is not clear what gives these wise men and women the power to take such consequential decisions and whether they can act independently from the platform that appointed them.[147]

Moderation always involves **trade-offs and value choices**. The risk is that by trusting or allowing private companies (or select groups of wise people) to make such choices, society loses 'an important opportunity to scrutinize norms and socially negotiate the values trade-offs they embed'.[148]

[141] F. Fukuyama, 'Making the Internet safe for democracy', *Journal of Democracy*, 32 (2), 2021.

[142] For example, TikTok was criticised for presumably instructing its content moderators to filter out people and spaces who are 'deviant'. See Gillespie, et al., 'Expanding the debate about content moderation', 2020.

[143] A. Satariano, 'After barring Trump, Facebook and Twitter face scrutiny about inaction abroad', *New York Times*, 2021.

[144] Fredheim et al., 'Social media manipulation report', 2020, p. 4.

[145] See, for example, A. Pasternack, 'How Facebook pressures its fact-checkers', *Fast Company*, 2020.

[146] Alaphilippe et al., 'Automated tackling of disinformation', 2019, p. 41.

[147] See, for example, F. Patel and L. Hecht-Felella, 'Facebook bylaws for takedown Oversight Board: Questions of independence', *Just Security*, 2020.

[148] Elkin-Koren, 'Contesting algorithms', 2020, p. 7.

Key points – Risks of moderation

- whereas moderation measures can help tackle disinformation on social media, they can also undermine freedom of expression and enable political censorship;
- whereas all moderation measures are risky, content removal is particularly problematic when targeted content is not explicitly illegal;
- deleting and labelling content can be counterproductive, as it may reinforce perceptions about unfair and unjustified censorship of particular views and groups;
- whereas automation can alleviate some burdens of human moderation, it can also amplify errors and automate pre-existing bias;
- increased pressure on social media to tackle problematic content may push platforms to rely even more strongly on automated tools, thus leading to over-censorship;
- despite efforts to make moderation more transparent and systematic, moderation measures on social media remain largely unclear, arbitrary, and inconsistently applied;
- social media moderation raises a serious problem of accountability: why should social media platforms decide what problematic (hence removable) content is?

6. Microtargeting

6.1. Social media and microtargeting

Microtargeting describes advertising strategies that use data and predictive modelling techniques to disseminate highly personalised messages to influence individual behaviour. Political microtargeting refers to a subset of such strategies aiming to influence citizens' political opinions and behaviour, i.e., to persuade or dissuade, inform or confuse, and mobilise or demobilise voters.[149]

Targeted advertising, including for political purposes, is nothing new. Political campaigners have for decades used demographic data (such as age, education, employment, or residence) to refine and focus electoral strategies (e.g. voter segmentation). The novelty of political microtargeting is the use of greater and **broader types of data** than conventional advertising,[150] including data on online behaviour (e.g. purchase history, browsing history, social media likes and shares). This allows targeters to **identify micro-groups** of people who share certain characteristics or inclinations that make them more likely to respond to a specific message. **Psychological profiles** may also be used to capture personal inclinations and predispositions. While appealing to people's unconscious fears and yearnings is not a novel advertising tactic, microtargeting provides far more effective ways to identify these predispositions and to reach people who have them. Lastly, unlike traditional advertising, which is able to adapt its messages to none other but the general demographic groups, microtargeting allows **messages to be tailored** to much smaller and homogenous groups. For example, this can be done by using 'A/B testing' techniques, which imply sending out slightly different versions of the same message to different population segments to test patterns in their responses.

Social media and other online platforms rely on a variety of data for profiling and targeting people, and on efficient ways to 'approach narrow groups who share a similar identity with customised

[149] T. Dobber, R. 'O Fathaigh and F. J. Zuiderveen Borgesius, 'The regulation of online political micro-targeting in Europe', *Internet Policy Review*, 8(4), 2019.

[150] ibid., pp. 11-12.

political messages'.[151] Many social media platforms sell advertising services that allow political campaigners to send targeted messages to users.

The use of microtargeting for political purposes became a public issue in 2017 when it was revealed that **Cambridge Analytica**, a marketing company, had created psychographic profiles of over 220 million Americans for the purpose of using them as targets of political ads.[152] Following this scandal, several platforms started imposing limits on political advertising. For example, in 2019, Twitter banned all political advertisements on the platform globally. It then established a mandatory Cause-Based Ad certification and review process requiring advertisers promoting content that educates, raises awareness, or supports social causes to go through a certification process. While Facebook did not ban political advertising on its platform, it improved transparency by creating an Ad Library containing a searchable collection of all ads running on its apps and services. However, there is evidence showing that social media platforms continue to uphold **different standards** on political advertising across regions and countries.[153]

6.2. Key risks of microtargeting

- **political manipulation** – covert and targeted political advertising diminishes citizens' capacities for democratic self-determination;
- **distortion of the electoral process** – targeted political advertising distorts the electoral process, challenges existing rules and alters electoral outcomes.

6.2.1. Political manipulation

Political microtargeting can affect citizens' engagement, political communication and electoral processes. On the one hand, political microtargeting can be used to **re-engage citizens** in politics, allowing parties to understand better what citizens think and need and to better 'reach voters with customized information that is relevant to them'.[154] On the other hand, selective political communication can enable politicians to use flexible and even contradictory messages that are disconnected from officially held ideological positions ('digital gerrymandering'[155]) to manipulate citizens' views and expectations. In a democracy, it is generally understood and accepted that political messages may seek to persuade or mobilise citizens by exaggerating or making claims based on a 'slim factual basis'.[156] Although some citizens may fail to spot inconsistences and rhetorical tricks used in political communication, political messages are usually open for wider public scrutiny.

The problem with political microtargeting is that it is both **intrusive and covert**. It is intrusive because it is based on a great deal of data about a person and because it uses techniques that may reveal **highly personal characteristics** (such as psychological inclinations). Political microtargeting is covert because users are not typically **aware of being targeted** by it. They are usually also not aware that a particular message it only targeting them (and a few others in their micro-group). The covert nature of political microtargeting makes people vulnerable to manipulation. Manipulation, which can be defined as the intentional and covert influencing of someone's decision-making 'by

[151] International Institute for Democracy and Electoral Assistance, 'Digital microtargeting: political party innovation primer 1', 2018, p. 10.

[152] H. Grassegger, and M. Krogerus, 'The data that turned the world upside down', *Vice*, 2017.

[153] Privacy International, 'Online political ads - a study of inequality in transparency standards', 2021.

[154] International Institute for Democracy and Electoral Assistance, 'Digital microtargeting', 2018, p. 7.

[155] J. Zitrain, 'Engineering an election: Digital gerrymandering poses a threat to democracy', *Harvard law Review Forum*, 2014.

[156] As the Court found in the case *Arbeiter v. Austria*, a politician's freedom of expression is protected even when their claims rely on a 'slim factual basis'. European Court, Case *Arbeiter v. Austria*, judgement of 25 January 2007.

targeting and exploiting their decision-making vulnerabilities',[157] undermines individual autonomy and hence citizens' capacity to form and make political choices.

The intrusive and covert nature of political microtargeting **compromises citizens' privacy**[158] and may **impede their capacity to form and express political judgements**. Political microtargeting can 'considerably limit freedom of information of potential voters, who would make their voting choices lacking impartial information on political candidates'.[159] For example, microtargeting techniques may be used to 'strategically disseminate disinformation to particularly susceptible groups of users'.[160] They may seek to exploit individuals' psychological vulnerabilities by pushing messages that are more likely to impress and persuade the person (e.g. a paranoid may receive messages that are fear-based). They may create psychological wants that masquerade as cognitive choices, thus eroding the foundational democratic principles of free will, equality, and fairness.[161] Although a lot of attention has been paid to the use of psychological profiling, the use of intrusive and non-transparent data-driven targeting for electoral campaigning poses a great challenge to democracy even if it does not rely on psychographics.[162]

6.2.1. Distortion of the electoral process

Political microtargeting practices challenge established **electoral rules** concerning transparency, campaigning and political funding. They tend to blur the boundary between permissible election campaigns and systematic voter manipulation and thus 'undermine the ability of existing regulation to maintain a level playing field in electoral communication'.[163]

Existing evidence suggests that political **microtargeting has been increasingly used** in recent electoral campaigns around the world. During the 2016 US presidential campaign, Cambridge Analytica, a marketing company, created psychographic profiles of over 220 million Americans to target political ads at them. According to estimates, Trump, then a candidate, ran 5.9 million ads to identify and then promote those ad versions that generated the greatest Facebook engagement.[164] In the UK, during the Brexit referendum, the VoteLeave campaign used targeted ads that contained false information about the high cost of Britain's EU membership and the imminent accession of Turkey to the EU.[165] Political microtargeting practices have been documented in the context of recent national elections in the UK, the Netherlands, Germany, and France.[166] For example, it is reported[167] that during the 2021 Dutch elections political parties used microtargeting techniques to target would-be supporters.

According to the 2021 EU citizens' **survey on democracy**,[168] more than half of citizens are concerned about: elections being manipulated through cyberattacks (57 %); foreign actors and

[157] D. Susser, B. Roessler, and H. Nissenbaum, 'Technology, autonomy, and manipulation', *Internet Policy Review*, 8 (2), 2019, p. 4.

[158] K. Manheim, and L. Kaplan, 'Artificial intelligence: Risks to privacy and democracy', *Yale Journal of Law & Technology*, 21, 2019, p. 25

[159] M. Brkan, 'EU fundamental rights and democracy implications of data-driven political campaigns', *Maastricht Journal of European and Comparative Law*, 27 (6), 2020, p. 776.

[160] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 35.

[161] Manheim and Kaplan, 'Artificial intelligence', 2019.

[162] Vaidhyanathan, *Antisocial media*, 2018, pp. 160-2.

[163] D. Tambini, S. Labo, E. Goodman, and M. Moore, 'The new political campaigning', *LSE policy brief*, 2027.

[164] Grassegger and Krogerus, 'The data that turned the world upside down', 2017.

[165] C. Cadwalladr, "Plucky little panel" that found the truth about fake news, Facebook and Brexit',*The Guardian*, 2018.

[166] Dobber et al., 'The regulation of online political micro-targeting in Europe', 2019.

[167] D. Beraldo, S. Milan, C. Agosti, B. N. Sotic, R. Vliegenthart, S. Kruikemeier, L. P. Otto, S. A. Vermeer, X. Chu, and F. Votta, 'Political advertising exposed: Tracking Facebook ads in the 2021 Dutch elections', *Internet Policy Review*, 2021.

[168] Special Eurobarometer 507: Democracy in the EU, report, 2021.

criminal groups influencing elections covertly (55 %); election results (53 %) being manipulated; about people being pressured into voting a particular way (52 %).

As in the case of disinformation, the actual impact of microtargeting **is hard to quantify**. The authors of a 2012 study,[169] carried out by a team that featured two senior data scientists at Facebook, claimed that targeted Facebook messages could greatly influence 'political self-expression, information seeking and real-world voting behaviour of millions of people, including Facebook users, as well as their ''friends, and friends of friends'''. Another paper,[170] published in 2013, showcased a predictive model that could 'automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits…' from Facebook Likes. In 2017, three studies by researchers from Cambridge University tested the effects of psychological persuasion on people's purchasing behaviour[171] using data from 3.7 million people who took a personality test via a Facebook app. The research showed that people who were shown ads that matched their personality were more likely to click on these ads and to purchase the advertised product than people who were served mismatched ads. The authors suggested that 'the application of psychological targeting makes it possible to influence the behaviour of large groups of people by tailoring persuasive appeals to the psychological needs of the target audiences'.[172]

Critics have taken **issue with this research** pointing to, for example, confusions between correlation and causation, and insufficient consideration of the selection effect of ad targeting (e.g. Facebook ads will be shown to people who are more likely to respond to them in the first place).[173] Doubts also exist about whether the psychological persuasion works as well in different contexts or for different issues, say for persuading people to change their political preferences (as opposed to buying stuff).

Even if political microtargeting is less effective than commercial microtargeting, it is possible that advertising strategies targeting the right group of people in the right moment and at the right place could potentially **change the result of an election**.[174]

The impact of microtargeting may depend on the **type of electoral systems**. In winner-takes-all systems (where the political party or group with the most votes gets all the seats within a given district), political microtargeting that focuses on small but key segments of society (e.g. 'swing voters') may be determine election results. For example, the use of political microtargeting by Cambridge Analytica has been credited as one of the key elements of Trump's electoral success in 2016. In hindsight, it appears that campaigners needed to persuade only a small fraction of voters to achieve the result, as it took about 80 000 votes in three key states to tip the scales.[175] Whereas this situation is less likely to occur in proportional electoral systems (where seats are distributed among parties according to the share of the votes received), even there 'a significant microtargeting campaign directed to the right sectors of society might be able to tip the balance of an election'.[176]

---

[169] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, 'A 61-million-person experiment in social influence and political mobilization', *Nature*, 489, 2012.

[170] M. Kosinski, D. Stillwell, and T. Graepel, 'Private traits and attributes are predictable from digital records of human behavior', *Proceedings of the national academy of sciences* 110 (15), 2013.

[171] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell, 'Psychological targeting as an effective approach to digital mass persuasion', *Proceedings of the national academy of sciences* 114 (48), 2017.

[172] Matz et al., 'Psychological targeting as an effective approach', 2017, p. 12714.

[173] Eckles, D., Gordon, B. R., & Johnson, G. A. (2018). Field studies of psychologically targeted ads face threats to internal validity. *Proceedings of the National Academy of Sciences*, *115*(23), E5254-E5255.

[174] Aral, *The hype machine*, 2021, p. 224.

[175] P. Bump, 'Donald Trump will be president thanks to 80,000 people in three states', *Washington Post*, 1 December 2016.

[176] International Institute for Democracy and Electoral Assistance, 'Digital microtargeting', 2018, p. 10.

Lastly, even if political microtargeting cannot be blamed for tipping recent elections, the concern here is that techniques based on this approach may become even **more effective in the future**, given the strong political and economic interests and the prospect of future technological improvements (e.g. advances in discursive AI algorithms making it possible to draft highly targeted messages quickly and on a large scale).

Key points – Risks of microtargeting

- compared to conventional political advertising, political microtargeting leverages broader types of data to identify micro-groups of people and to test and serve them tailored messages;
- whereas microtargeting can serve to re-engage citizens in politics, it can also be used to manipulate citizens' views and expectations;
- microtargeting is intrusive because it is based on a great deal of data and insights (including psychological profiles) about people, which can undermine privacy and autonomy;
- the covert or hidden nature of microtargeting increases the risk of manipulation and thus undercuts citizens' capacity to form and make political choices;
- political microtargeting challenges existing electoral rules concerning transparency, campaigning and political funding and can distort electoral and political processes;
- whereas the evidence about the wide use of political micro-targeting is growing, the actual impact of political microtargeting remains uncertain;
- given the nature of political competition, it is possible that political microtargeting campaigns can determine the outcome of elections, in particular in winner-takes-all electoral systems;
- even if microtargeting cannot be blamed for tipping recent elections, the risks it creates are likely to increase, given the high political and economic interests at stake and future technological advances.

7. Key policy approaches

The risks to democracy discussed in this paper are generated through a **complex interaction of actors** (social media platforms, advertisers, regulators, political parties, citizens, etc.)[177] whose intended or non-intended actions are shaped by various economic, technological, and normative systems. Tackling these risks will require **measures in many areas,** including competition, data protection, electoral law, technological design, research, education and citizen engagement. It will also require measures at **national, EU and international/global level**.

The EU already has laws and policies in place to tackle many of the risks associated with social media (e.g. strong data protection rules) and is spearheading efforts to counteract these risks. Several **key approaches** to tackling social media risks to democracy can be identified:

- **enhance competition** to combat abuse of market dominance;
- **protect data and privacy** to prevent abusive processing of data;
- **review content liability rules** to clarify responsibilities for online content;
- **increase transparency and accountability** for filtering and moderating content;
- **oversee algorithms** to increase their transparency and reliability;
- **regulate targeted political advertising** to prevent abuse and manipulation;
- **empower citizens** to enable them to understand digital risks and fend off attacks.

[177] Bertolini et al., 'Liability of online platforms', 2021, p.72.

7.1. Enhance competition

There are growing concerns that the **market power** accumulated by several online platforms, including social media platforms, damages competition in the digital market and leads to abuses of users' rights. For example, by operating in different markets, major platforms can integrate behavioural data from various services and advertising networks to the detriment of privacy, consumer rights, and competition.[178] Moreover, because major online platforms benefit from the fact that users are more likely to choose platforms with a large user base (strong **network effects**), they have little incentives to allow or enable users to switch to competitive platforms.

The EU **competition rules** allow the Commission to investigate and sanction abuses of dominant market position, including by imposing hefty fines. Merger control measures can be used to regulate market conduct as a way to prevent anti-competitive effects that are obtained, among others, through the accumulation of data from various sources. However, such measures are not always easy to implement. For example, competition authorities often have difficulties proving anti-competitive effects, given the limited information available to them and the high uncertainty.[179] To counter anti-competitive behaviour by online platforms, the EU has adopted a series of new rules. For example, the Platform-to-Business Regulation, which entered into force in July 2020, provides for new transparency rules and redress mechanisms for businesses using online platforms' services.

The proposed **digital markets act** (DMA)[180] seeks to establish a regulatory framework for major digital platforms designated as gatekeepers. It imposes new obligations on gatekeepers (e.g. data access obligation) and prohibitions (e.g. ban on self-preferencing). For example, it requires a gatekeeper to refrain from combining personal data obtained from its core platform services with personal data from other services offered by the gatekeeper or third parties, unless the end user provides their consent.

Another way of reducing the power of online platforms is to impose stricter obligations regarding **data portability and interopability**.[181] The high costs of switching between platforms gives major platforms an advantage and leaves users with few choices. Reducing users' costs of switching between platforms may increase competition in digital markets and force platforms to pay more attention to users' rights and views.[182] Whereas interoperability can reduce entry barriers associated network effects, it may also generate unintentional competition harms, for example, by imposing significant burdens on new entrants and entrenching incumbents' systems and technologies.[183]

The **General Data Protection Regulation** (GDPR) provides for the right of data subjects to obtain personal data they have provided to a data controller in a structured, commonly used and machine-readable format as well as to transfer such data to a different controller (Article 20). Despite its potential, the implementation of the right to data portability has been affected by a lack of clarity regarding the object of the right and its relationship with other rights.[184] For example, it is disputed

[178] G. Sartor, The impact of the General Data Protection Regulation (GDPR) on artificial intelligence, European Parliament, EPRS, 2020, p. 19.

[179] J-U, Franck, M. Peitz, 'How to challenge big tech', *VerfBlog*, 6 September 2021.

[180] European Commission, Proposal for a Regulation on contestable and fair markets in the digital sector (Digital Markets Act), COM/2020/842 final. For an analysis, see T. Madiega, Digital markets act, EPRS, European Parliament, 2021.

[181] Data portability refers to the right to obtain and take your data from one online service to another. Interoperability refers to the ability of different computerized products or services to interact with one another and work seamlessly.

[182] C. Doctorow, 'Competitive Compatibility: Let's Fix the Internet, Not the Tech Giants', *Communications of the ACM*, 64 (10), 2021.

[183] OECD, 'Data portability, interoperability and digital platform competition', OECD Competition Committee Discussion Paper, 2021.

[184] P. De Hert, V. Papakonstantinou, G. Malgieri, L. Beslay, and I. Sanchez, 'The right to data portability in the GDPR: Towards user-centric interoperability of digital services', *Computer Law & Security Review*, 34 (2), 2018.

whether the right covers only personal data that are 'provided' by the person or also includes behavioural data or inferences about the user.

The **DMA proposal** introduces broader data portability requirements to ensure additional forms of portability, including the portability of non-personal data for business users and real-time and continuous portability. In its opinion on the DMA proposal, the European Data Protection Supervisor (EDPS) stated that 'increased interoperability has the potential to facilitate the development of a more open, pluralistic digital environment' and recommended 'introducing minimum interoperability requirements for gatekeepers, with explicit obligations on gatekeepers to support interoperability, as well as obligations not to take measures that impede such interoperability'.[185]

7.2. Protect data and privacy

The market power of online platforms depends greatly on their ability to acquire and retain users, which often relies on extracting and analysing as much data as possible (including personal and behavioural data of users and non-users). This **data extractive imperative** generates risks of data protection and privacy, which may weaken citizens' capacity for democratic self-determination. Establishing and enforcing strong data protection and digital privacy rules can reduce such risks. For example, providing users with more control over their data and requiring online platforms to observe strict data protection rules should weed out abusive practices, such as unlawful profiling, discriminatory targeting and manipulation.

The EU has developed a strong framework for data protection and digital privacy. The **GDPR** provides strong legal safeguards against digital surveillance, including a set of rights of data subjects and obligations of data controllers on acquiring consent, providing information and refraining from profiling. The **ePrivacy Directive**[186] (currently under revision) provides rules on the confidentiality of communications, tracking and monitoring. For example, it requires users' consent for cookies and other tracking devices that interfere with the users' terminal equipment.

Despite their ground-breaking potential, the **enforcement of the EU data protection and privacy** rules remains challenging. For example, the consent-based approach promoted by the GDPR has been challenged by the fact that, given the presence of great information asymmetries between service providers and consumers, users have 'limited ability to successfully exercise their rights when interacting with digital services'.[187] Some rules are not sufficiently clear. For example, the GDPR does not straightforwardly exclude the possibility that consent to the processing of personal data can be free when this is a condition for the provision of a service; as a result, this could lead to 'unlawful or borderline practices ... through which users are induced to consent to all kinds of processing of their data'.[188] **The use of AI** creates additional challenges for the interpretation of EU data protection principles such as purpose limitation, data minimisation, the special treatment of sensitive data, and the limitation on automated decision data. Whereas the GDPR can be interpreted and applied in such a way that it does not substantially hinder the application of AI to personal data, there may be a number of AI-related data-protection issues do not have an explicit answer in the GDPR.[189]

Data protection safeguards are also **a prerequisite for fair and democratic elections**.[190] Whereas the organisation of elections in the EU is largely regulated at Member State level, the EU rules on the

[185] EDPS, Opinion on the Proposal for a Digital Markets Act, 2021.

[186] Directive 2002/58 concerning the processing of personal data and the protection of privacy in the electronic communications sector.

[187] A. Koene, C. Clifton, Y. Hatada, H. Webb, M. Patel, C. Machado, J. LaViolette, R. Richardson, and D. Reisman, A governance framework for algorithmic accountability and transparency, EPRS, European Parliament, 2019, p. 81.

[188] G. Sartor, Regulating targeted and behavioural advertising in digital services: How to ensure users' informed consent, Directorate-General for Internal Policies, European Parliament, 2021, p. 12.

[189] Sartor, The impact of the General Data Protection Regulation (GDPR) on artificial intelligence, 2020, p. 7.

[190] EDPS, 'Europe votes 2019: Data protection is a prerequisite for fair and democratic elections', 16 May 2019.

processing of personal data are fully applicable to both European and national political parties and other actors in the electoral context, such as data brokers and social media platforms. The GDPR provides for special protection of sensitive data, including people's political opinions. A 2019 amendment of the Regulation on the statute and funding of European political parties and European political foundations introduced a verification procedure and sanctions related to infringements of rules on the protection of personal data by a European political party or a European political foundation in the context of elections to the European Parliament.

Whereas data protection and digital privacy are essential for addressing social media risks, there are **limits to how much can be achieved** within the existing data protection framework. For example, the GDPR does not contain specific rules that directly address targeted advertising (microtargeting), although its provisions can be interpreted as addressing this issue. For example, the European Data Protection board (EDPB) has issued guidelines on the targeting of social media users.[191] However, it may be possible to develop sophisticated advertising tools and strategies that comply with data protection rules (e.g. explicit consent) and yet manipulate voters.[192]

7.3. Review content liability rules

Social media platforms provide accessible and effective ways to spread content to wide audiences, including content that is illegal, harmful, misleading, and/or inaccurate. Amid concerns about the widespread disinformation on social media, there are calls to **review the legal responsibilities** of these platforms regarding the content they host and promote.[193] According to a recent report,[194] by March 2020, at least 28 countries had passed laws to address online disinformation involving, among others, revisions of electoral laws, cybersecurity and penal codes.

The EU has a complex **legal framework on content liability** rules comprised of both hard-law rules at both EU and national level, and voluntary instruments such as codes of conduct.[195] For example, online platforms can be obliged to swiftly remove illegal content such as hate speech, content infringing copyrights and terrorist content. The E-Commerce Directive (Directive 2000/31) grants liability protection to providers of information society services in relation to user-generated content. Hosting providers benefit from a liability exemption provided they act expeditiously to remove or disable access to information upon obtaining knowledge about its illegal character (Article 14). Intermediaries can be ordered by competent authorities to terminate or prevent infringements by their users, but they cannot be obliged to generally monitor content or to 'actively seek facts or circumstances indicating illegal activity' (Article 15).

To be considered an intermediary, a service provider must be neutrally providing a service by automatic, technical and passive means. This means that service providers no longer act as intermediaries where they take an active role in relation to content that would give them either knowledge of or control over that content. Generally, liability increases as the intermediary's editorial control increases.[196] In the context of efforts to re-examine the responsibilities of online platforms for content, a key question is whether certain platforms should be considered mere intermediates. Some argue that **recommender systems**, which actively promote certain content,

[191] EDPB, 'Guidelines 8/2020 on the targeting of social media users', 2 September 2020.

[192] Dobber et al., 'The regulation of online political micro-targeting in Europe', 2019.

[193] For an overview, see T. Madiega, Reform of the EU liability regime for online intermediaries: Background on the forthcoming digital services act, EPRS, European Parliament, April 2020.

[194] Broadband Commission for Sustainable Development, Balancing act: Countering digital disinformation while respecting freedom of expression, International Telecommunication Union (ITU) and the United Nations Educational, Scientific and Cultural Organization (Unesco), 2020.

[195] Bertolini et al., 'Liability of online platforms', 2021.

[196] Marsden and Meyer, 'Regulating disinformation with artificial intelligence', 2019. p. 23.

cannot be considered intermediary service providers and thus extending liability protection to them should 'come with some responsibilities beyond simply removing illegal content expeditiously'.[197]

In one of its resolutions[198] with recommendations for a digital services act (DSA) proposal, the European Parliament suggested that 'online platforms should place effective and appropriate safeguards, in particular to ensure that they act in a diligent, proportionate and non-discriminatory manner, and to **prevent the unintended removal of content which is not illegal**'. The Parliament urged establishing a clear legal framework for the removal of illegal content without imposing general monitoring obligations on digital service providers. It called for a 'strict distinction to be made between illegal content, punishable acts and illegally shared content on the one hand, and harmful content, hate speech and disinformation on the other, which are not always illegal'.

The proposed DSA[199] introduces new due diligence obligations for online intermediaries but preserves the current liability exemptions. It also provides a 'Good Samaritan's rule' allowing intermediaries to take steps to detect, identify and remove/disable access to illegal content while still benefiting from liability exemptions (Article 6).

7.4. Increase transparency and accountability

Whereas many social media platforms have recently tightened their internal rules and standards to tackle problematic content and abusive practices, concerns remain about the transparency, consistency and accountability of these measures. To address these issues, the Commission established an **EU Code of Practice on Disinformation**,[200] through which the participating platforms committed to work towards: reducing the economic incentives for the dissemination of disinformation online; enhancing the transparency of political advertising; tackling manipulative techniques; prioritising trustworthy information; and engaging in collaborative activities with fact-checkers.[201] However, this self-regulatory approach has so far produced limited results.

A major issue is **partial and uneven compliance**.[202] For example, a 2021 report by the European Court of Auditors (ECA) found that disinformation in the EU is 'tackled, but not tamed'[203] and that the code has not been able to hold online platforms to account for their actions against disinformation. Following its 2021 assessment of the implementation of the code,[204] the Commission expressed its intention to overhaul it into a co-regulatory framework. It called upon signatories to make 'stronger and more specific commitments' and invited other platforms to join. The Commission issued guidance[205] on strengthening the code by establishing a more robust monitoring framework, empowering users, encouraging better cooperation between the platforms and fact-checkers, and providing a framework for access to data for researchers.

In September 2020, the European Parliament set up a **temporary special committee** on foreign interference in all democratic processes in the EU, including disinformation (INGE). In a 2020 working

[197] Cobbe and Jatinder, 'Regulating recommending motivations, considerations, and principles', 2019.

[198] European Parliament, Resolution of 20 October 2020 with recommendations to the Commission on the Digital Services Act: Improving the functioning of the Single Market.

[199] European Commission, Proposal for a Regulation on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, 2020.

[200] European Commission, Code of Practice on Disinformation.

[201] Other measures include the establishment of the EEAS task force, the launch of the European Digital Media Observatory, and funding of research and innovation projects tackling disinformation.

[202] Bertolini et al., 'Liability of online platforms', 2021, p. 74.

[203] ECA, Disinformation affecting the EU: tackled but not tamed, special report, September 2021.

[204] European Commission, Assessment of the Code of Practice on Disinformation – Achievements and areas for further improvement, Staff Working Document, 2020.

[205] European Commission, Guidance on Strengthening the Code of Practice on Disinformation, 2021.

document,[206] the INGE rapporteur called for developing 'stricter rules for regulating platforms with regard to transparency, sanctions, the duty to provide linguistic expertise and cooperate across platforms, as well as clear boundaries to prevent abuse of users' data'.

The **proposed DSA introduces new transparency and accountability obligations** for providers of intermediary services, and in particular online platforms, such as social media ones,[207] requiring them to clearly indicate in their terms and conditions 'information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review'. Moreover, they 'shall act in a diligent, objective, and proportionate manner in applying and enforcing the restrictions ... with due regard to the rights and legitimate interests of all parties involved, including the applicable fundamental rights of the recipients of the service as enshrined in the Charter' (Article 12). Very large online platforms (VLOPs) are required to abide to a higher standard of transparency and accountability, for example, regarding moderation decisions, advertising and algorithmic processes.

Another key issue is that the decisions taken by **social media platforms are not accountable**, hence widespread calls for the establishment of independent appeal and audit mechanisms to improve oversight and accountability of these online platforms.[208] The **proposed DSA** does not prohibit platforms from introducing their own internal rules, but it obliges them to establish clear and unambiguous rules and to apply them in a proportionate manner. Platforms will have to put in place **notice and action mechanisms** allowing individuals or entities to notify platforms of the presence of content that they consider to be illegal (Article 14). They will have to inform users about the reasons for blocking them (Article 15) and to establish an effective **internal complaint-handling system** enabling users to challenge the platforms' decisions (Article 17). VLOPs will be subject to regular and **independent auditing** to assess for their compliance with the obligations (Article 28). They will also need to appoint compliance officers (Article 32) and submit regular transparency reports (Article 33). VLOPs will be required to **carry out a risk assessment** identifying any significant systemic risks stemming from the functioning and use of their services and to implement reasonable, proportionate and effective mitigation measures (Articles 26 and 27). These systemic risks are quite broadly conceived and include intentional manipulations of services with a 'foreseeable negative effect' on civic discourse or with effects on 'electoral processes and public security'.[209] VLOPs will also have to provide researchers with **access to data** as would enable them to conduct research that contributes to the identification and understanding of systemic risks (Article 31). Access to data is limited to 'vetted researchers' (with university affiliation) and to research on 'systemic risks', without jeopardising data security and the protection of confidential information including trade secrets.

As the DSA proposal tackles a great number of complex issues, the legislative work requires **balancing different concerns and interests**. For example, some critics take an issue with the **overall approach** of the DSA, arguing that it could lead to either insufficient regulation – where platforms assess that the risk obligations are too vague to require actions that are not already covered by internal compliance rules – or to over-regulation – where platforms excessively delete content 'associated with systemic risk-potential'.[210] There are concerns about the limitations on **access to data**. In August 2021, more than 50 civil society organisations and disinformation experts

[206] Working document on the state of the foreign interference in the European Union, including disinformation, Special Committee on Foreign Interference in all Democratic Processes in the European Union, including Disinformation, European Parliament, 17 December 2020.

[207] For an overview, see Madiega, Digital services act, 2021.

[208] See, for example, Marsden and Meyer, 'Regulating disinformation with artificial intelligence', 2019. p. 6.

[209] For a critical discussion see, A. Peukert, 'Five reasons to be skeptical about the DSA', *VerfBlog*, 2021.

[210] H. Ruschemeier, 'Re-subjecting state-like actors to the state: Potential for improvement in the Digital Services Act', *VerfBlog*, 2021.

signed a letter[211] requesting the EU legislators to strengthen the **accountability** provisions of the proposed DSA. They requested, among other things, to expand access to platform data to civil society researchers and journalists and to establish an independent European Oversight Board to oversee implementation of the DSA. Others have pointed out that the DSA proposal does not address the issues of **independent 'scraping'** of data from platforms[212] – a practice involving the use of data found on platforms to fact-check the official data provided by platforms.[213] The risk is that platforms would continue to 'weaponise' their terms of services 'against individuals or organisations that attempt to hold large platforms to account'.[214]

7.5. Oversee algorithms

There are two types of algorithmic challenges related to social media platforms. First, the underlying algorithms of these platforms may be designed and/or operate in ways that are undermining fundamental rights, individual autonomy and democracy (e.g. amplifying disinformation, enabling manipulation, and fostering social and political polarisation). Second, external parties may use automated tools (e.g. bots) to abuse and manipulate underlying algorithms to pursue goals that, directly or indirectly, undermine fundamental rights, individual autonomy and democracy. Concerns about the negative impact of bots have led to calls to curtail the use of bots for automated amplification of disinformation[215] and to ban bots from disseminating political and public issue ads.[216] However, removing bots and tackling abusive behaviour typically require deploying more automated tools, which also increases the risks of over-filtering and censorship.[217]

A key issue with social media algorithms (and algorithms in general) is that they are **black boxes**, meaning that their inner workings are both difficult to explain and kept secret to protect intellectual property rights and to prevent tempering. There are several suggestions about how to open black box algorithms. Firstly, social media platforms can be required to provide **more information to users** about how their algorithms (e.g. news recommendation algorithms) work. However, algorithmic transparency[218] and explainability are complex and evolving concepts, and it is not clear whether they could be achieved through general transparency requirements. Secondly, algorithms should be subject to **independent auditing** to ensure that they do not violate fundamental rights. Another common suggestion is to ensure adequate **human oversight** of algorithms. However, there are concerns that humans may not be able to oversee algorithms and that therefore an oversight mechanism may provide a false sense of security and 'legitimize government use of flawed and controversial algorithms without addressing the fundamental issues'.[219]

Research has shown that many negative effects of social media are the result of a **complex interaction between machines and humans**. However, more knowledge and understanding about this interaction is needed. Apart from informing users and establishing a narrow audit mechanism, this would require **broader scrutiny** from various stakeholders[220] and **wider access to**

[211] EU Disinfo Lab, 'Open letter to EU policy-makers: How the Digital Services Act (DSA) can tackle disinformation', 31 August 2021.

[212] There were a series of incidents recently where platforms blocked access to data or tried to intimidate independent researchers who were studying the network. See, for example, L. Edelson and D. McCoy, 'Facebook is obstructing our work on disinformation. Other researchers could be next', *The Guardian*, 14 August 2021; Algorithm Watch, 'AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook', 13 August 2021.

[213] See, for example, NYU Ad Observatory, 'The Political Ads Facebook Won't Show You'.

[214] Algorithmic Watch, 'Under Facebook's thumb: Platforms must stop suppressing public interest research', 2021.

[215] Alaphilippe et al., 'Automated tackling of disinformation', 2019.

[216] Bayer et al., 'Disinformation and propaganda – impact on the functioning of the rule of law', 2019.

[217] Marsden and Meyer, 'Regulating disinformation with artificial intelligence', 2019. p. 6.

[218] Koene et al., 'A governance framework for algorithmic accountability', 2019, pp. 17-18.

[219] B. Green, 'The flaws of policies requiring human oversight of government algorithms', SSRN, 2021.

[220] A. Alaphilippe *et al.*, 'Automated tackling of disinformation', 2019, p. 5.

data. Despite efforts by social media platforms to cooperate with researchers, it is still difficult for civil society stakeholders to obtain meaningful information on how platforms act to counter disinformation.[221]

In its 2020 resolution on DSA,[222] the Parliament stressed that algorithms used in digital services need to fully comply with requirements on fundamental rights and called on the Commission to introduce **transparency and accountability** requirements regarding automated decision-making processes, while ensuring compliance with requirements on user privacy and trade secrets.

A key issue is about who oversees algorithms. Options include private oversight (self-regulation) and independent oversight via law enforcement mechanisms or broad civil society and academic scrutiny (for which access to data is essential). The DSA obliges VLOPs to provide **access to data** to 'vetted researchers' to study issues related to systemic risks (Article 31). Furthermore, when using **recommender systems**, VLOPs must set out in their terms and conditions, 'in a clear, accessible manner and easily comprehensible manner', the main parameters used in their recommender systems, and any options to allow users to modify these parameters, including at least one option that is not based on the profiling of users (Article 29). The concern is that this may not be sufficient to protect users, since, as EDPS puts it, 'including information about the recommender system parameters and options in the terms and conditions would only make them difficult to find and understand for data subjects'.[223] Rather than empowering users, such provisions may create a fake sense of transparency.[224]

Finally, critics have pointed out that focusing too much on algorithms and their alleged effects (e.g. filter bubbles and polarisation) may be misguided. There is the risk of turning algorithms into a **technological scapegoat**[225] when, in fact, many of the issues at stake have deeper social and economic causes, such as rising inequality and public dissatisfaction with political institutions.[226] If these critics are right, solving these issues would require broader and more complex solutions than technological 'fixes'.[227]

7.6. Regulate targeted political advertising

There are several measures available to protect elections and democratic process from the risks posed by social media platforms. These include general measures, such as on data protection and general transparency, and specific measures to regulate targeted political advertising.

Data protection is a necessary yet not a sufficient prerequisite for addressing challenges of targeted political advertising.[228] Election law depends to a large degree on **transparency, both in terms of funding and electioneering**. The challenge as regards the online environment is that electoral rules are either not fully applicable or very hard to enforce. For example, political content online cannot be easily distinguished from other content and can be circulated outside of electoral and campaign periods.[229] In the EU, another challenge is that Member States have their own rules on political campaigning and advertising.

[221] EU Disinfo Lab, 'Overcoming obstacles to Exposing Disinformation', 2021.

[222] European Parliament, Digital Services Act: Improving the functioning of the Single Market, 20 October 2020.

[223] EDPS, Opinion on the Proposal for a Digital Services Act.

[224] N. Helberger, M. van Drunen, S. Vrijenhoek, J Möller, 'Regulation of news recommenders in the Digital Services Act: empowering David against the Very Large Online Goliath', *Internet Policy Review*, 2021.

[225] A. Bruns, 'It's not the technology, stupid', 2019, p. 3.

[226] F. Zimmermann, and M. Kohring, 'Mistrust, disinforming news, and vote choice', 2020.

[227] D. Frau-Meigs, 'Societal costs of "fake news"', 2018.

[228] M. Brkan, 'EU fundamental rights and democracy implications of data-driven', 2020, p. 778.

[229] J. Bayer *et al*., 'Disinformation and propaganda – impact on the functioning of the rule of law', 2019.

According to several studies, there is a need to **increase transparency requirements regarding online political advertising and campaign financing**. For example, it is suggested to extend transparency rules from offline to online advertising, while also adequately labelling political ads online, including content from influencers,[230] to allow voters to identify them as such.[231] Furthermore, consumers should be given more information about the existence of online targeted advertising, consent mechanisms should be strengthened and discrimination should be prevented.[232] Other suggestions focus on imposing stricter disclosure obligations on platforms, including requiring them to maintain a searchable repository of active and historical political and issue-based advertising targeting persons in the EU, and to disclose detailed information on campaign spending indicating, for example, the contracting media partner, the nature of the media content, and the targeting criteria. It is also suggested to create a regulatory body for political advertising or to task the European Public Prosecutor's Office with monitoring party expenditures.[233]

In its 2018 resolution[234] on the use of **Facebook users' data by Cambridge Analytica**, the European Parliament called on the Member States to introduce an obligatory system of digital imprints for electronic campaigning and advertising. It urged online platforms to ensure full compliance with the GDPR, distinguish political uses of their online advertising products from their commercial uses, and implement transparency features in relation to political advertising. It also called for a ban on 'profiling for political and electoral purposes and profiling based on online behaviour that may reveal political preferences. In one of its 2020 resolutions[235] on the DSA proposal, the European Parliament urged to **regulate targeted advertising** more strictly so as to ensure adequate consent, clear identification of paid advertisements or paid placement of sponsored content, and public scrutiny of ads (hosting platforms to maintain a publically accessible advertising archive indicating who has paid for them, and, if applicable, on behalf of whom). In another resolution,[236] the Parliament noted 'the potential negative impact of personalised advertising, in particular micro-targeted and behavioural advertisement' and called on the Commission to introduce additional rules on targeted advertising and micro-targeting and to consider introducing legislative measures to make online advertising more transparent.

The proposed DSA imposes **information requirements** for targeted advertising and additional transparency obligations on VLOPs with regard to ads repositories and recommendation systems. It obliges online platforms that display advertising on their online interfaces to ensure that each recipient can **identify advertisements** 'in a clear and unambiguous manner and in real time', as well as the natural or legal person on whose behalf the advertisement is displayed, and receive '**meaningful information** about the main parameters used to determine the recipient to whom the advertisement is displayed' (Article 24). Online platforms are required to make information about online advertisement publicly available until one year after the ad was displayed for the last time (Article 30). The mandatory **risk assessment** for VLOPs would allow to identify and mitigate systemic risks, including risks related to the intentional manipulation of their services that has a negative effect on civic discourse or on **electoral processes** and public security. VLOPs would have to take into account how their content moderation systems, recommender systems and systems for selecting and displaying advertisement influence these systemic risks.

---

[230] J. Bayer *et al*., 'Disinformation and propaganda – impact on the functioning of the rule of law', 2019.

[231] A. Alaphilippe *et al*., 'Automated tackling of disinformation', 2019.

[232] N. Fourberg, S. Taş, I. Wiewiorra, I Godlovitch, A. de Streel, H. Jacquemin, C. Bourguigon, F. Jacques, M. Ledger, M. Iognoul, J. Hill, M. Nunu, Online advertising: the impact of targeted advertising on advertisers, market access and consumer choice, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2021.

[233] J. Bayer *et al*., 'Disinformation and propaganda – impact on the functioning of the rule of law', 2019, p. 13.

[234] European Parliament, Resolution on the use of Facebook users' data by Cambridge Analytica and the impact on data protection, 25 October 2018.

[235] European Parliament, Digital Services Act: adapting commercial and civil law rules for commercial entities operating online, 20 October 2020.

[236] European Parliament, Digital Services Act: Improving the functioning of the Single Market, 20 October 2020.

In his opinion[237] on the DSA, the EDPS urged the co-legislators to consider additional rules going beyond transparency, including 'a phase-out leading to a **prohibition of targeted advertising** on the basis of pervasive tracking, as well as restrictions in relation to the categories of data that can be processed for targeting purposes and the categories of data that may be disclosed to advertisers or third parties to enable or facilitate targeted advertising.'

The EU Code of Practice on Disinformation provides that the platforms should ensure **transparency** about political and issue-based advertising and enable users to understand why they have been targeted by an advertisement. In its 2020 **European democracy action plan**,[238] the Commission recommended to Member States to focus on promoting the transparency of online political advertising, including campaign expenditure. It also announced the **review of the Regulation on the statute and funding of European political parties** and European political foundations with a view to: addressing the financing of European political parties from outside the EU; revising the audit requirements; strengthening the links between European financing and national campaigns; and facilitating transparency and auditing.

As part of a package of measures aimed at protecting election integrity and open democratic debate, in November 2021, the Commission presented a proposal[239] for a new regulation on transparency and targeting of political advertising. The proposal seeks to impose strict transparency requirements (labelling) for paid political advertising, which includes ads by, for or on behalf of a political actor as well as issue-based ads that are liable to influence the outcome of an election or referendum, a legislative or regulatory process or voting behaviour. It also introduces ban political targeting and amplification techniques that use or infer sensitive personal data, such as ethnic origin, religious beliefs or sexual orientation.

7.7. Empower citizens

Notwithstanding recent concerns about online platforms, **democracy has been under stress** long before the advent of the internet and social media. There is an acknowledged research field evidencing various ailments of established democracies, such as low voter turnout, rising populist political movements and widespread discontent with public institutions.[240] For a while, the internet and social media were seen as means to reinvigorate democracy by undercutting traditional gatekeepers and allowing (more) citizens to access information and express their views.

Although social media issues, such as disinformation, censorship and manipulation may not be solely responsible for the current flaws of democracies, they contribute to the **deterioration of trust** in information and institutions.[241] Even when these issues are not as big as they seem, exaggerated public perceptions of widespread disinformation and manipulation contribute to further demoralising and undermining citizens' trust in media and institutions.

Tackling the social media risk to democracy cannot be possible without engaging the main actors in a democracy: the citizens. There are several suggestions about how to engage and support citizens. First, citizens could benefit from more **accurate and timely information** about the content they see on social media, and about how their data is used. There is a growing number of tools available to verify and label online content either employed by social media platforms directly or promoted by independent fact-checkers. It is true that the effectiveness of correcting political

[237] EDPS, Opinion on the Proposal for a Digital Services Act, 2021.

[238] European Commission, Communication on the European Democracy Action Plan, COM/2020/790 final, 2020.

[239] European Commission, Proposal for a Regulation on the transparency and targeting of political advertising, 25 November 2021.

[240] See, for example, P. Mair, *Ruling the void: The hollowing of Western democracy*, Verso, 2013; Foa and Mounk, 'The danger of deconsolidation', 2016; Mounk, *The people versus democracy*, 2018.

[241] Stark and Stegmann, 'Are algorithms a threat to democracy', 2020, p. 37.

disinformation is contested. For example, debunking efforts cannot realistically match the scale of disinformation; moreover, debunking does not always reach the most vulnerable users and can have counter-productive effects (backlash effects). Nevertheless, there is research showing that more **user-centred approaches** that prompt users to consider the accuracy of social media content can improve the quality of their shared content.[242] Such proactive measures may be more effective because they appeal to users' individual responsibility.[243]

Broader measures on improving citizens' **media and digital literacy** could help to prevent harmful media influences and to promote a critical attitude to media consumption.[244] This could be done by establishing programmes on media literacy and on strengthening civic education on EU values of democracy and human rights at all levels.[245] Efforts should also be dedicated to **supporting institutions** that help citizens gain 'enlightened understanding' of public matters, such as independent media.

The EU has supported the creation of the Social Observatory for Disinformation and Social Media Analysis (SOMA) – bringing together researchers, fact-checkers, and media organisations – and the European Digital Media Observatory (EDMO), which facilitates coordination between fact-checking organisations, the scientific community, media practitioners and teachers, on the one hand, and technological platforms and public authorities, on the other. As announced in its 2020 European democracy action plan, the Commission is preparing a series of measures to further support **media pluralism** and to strengthen transparency of media ownership and state advertising.

8. Main references

Boehme-Neßler V., Digitising democracy: On reinventing democracy in the digital era - a legal, political and psychological perspective, Springer, 2020.

Bradshaw S., and Howard P. N., 'The global disinformation order: 2019 global inventory of organised social media manipulation', Oxford Internet Institute, 2019.

Brkan M, 'Artificial intelligence and democracy: The impact of disinformation, social bots and political targeting', *Delphi*, 2, 2019.

Dahl R., *On Democracy*, (first published by Yale University Press, 1998), Veritas, 2020.

Fredheim R., Bay S., Dek A., and Dek I., 'Social Media Manipulation Report 2020,' NATO Strategic Communications Centre of Excellence, 2020.

Grafanaki S., 'Autonomy challenges in the age of big data', *Fordham Intellectual Property, Media and Entertainment Law Journal*, 27 (4), 2016.

Newman N., Richard F., Schulz A., Andi S., and Kleis Nielsen R., 'Digital news report 2020', Reuters Institute and University of Oxford, 2020, pp. 18-19.

Stark B., and Stegmann D., 'Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse', Algorithm Watch, 2020.

Soroush V, Roy D., and Aral S., 'The spread of true and false news online', *Science*, 359, 2018.

Twetman H., Paramonova M., and Hanley M., 'Social Media Monitoring: A Primer', NATO Strategic Communications Centre of Excellence, 2021.

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

[242] Pennycook, G., Epstein, Z., Mosleh, M. *et al.* 'Shifting attention to accuracy can reduce misinformation online', *Nature*, 592, 2021.

[243] Müller and Denner, 'What can be done to counter Fake News', 2019.

[244] Alaphilippe et al., 'Automated tackling of disinformation', 2019, p. 5.

[245] Bayer et al., 'Disinformation and propaganda – impact on the functioning of the rule of law', 2019.

Wardle C., and Derakhshan H., 'Information disorder: Toward an interdisciplinary framework for research and policy making', Council of Europe, 2017.

9. Relevant EP studies and analyses

Alaphilippe A., Gizikis A., Hanot C., and Bontcheva K., 'Automated tackling of disinformation: Major challenges ahead', EPRS, European Parliament, 2019.

Bayer J., Bitiukova N., Bard P., Szakács J., Alemanno A. and Uszkiewicz P. E., 'Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States', Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, 2019.

Bentzen N., Trump's disinformation 'magaphone': Consequences, first lessons and outlook, EPRS, European Parliament, 2019.

Bertolini A., Episcopo H. and Cherciu N. A., Liability of online platforms, EPRS, European Parliament, 2021.

Bird E., Fox-Skelly J., Jenner N., Larbey R., Weitkamp E., and Winfield A., The ethics of artificial intelligence: Issues and initiatives, EPRS, European Parliament, 2020.

Colomina C., Sánchez Margalef H., and Youngs R., The impact of disinformation on democratic processes and human rights in the world, Policy Department for External Relations, European Parliament, 2021.

Fourberg N., Taş S., Wiewiorra I., Godlovitch I., de Streel A., Jacquemin H., Bourguigon C., Jacques F., Ledger M., lognoul M., Hill J., and Nunu M., Online advertising: the impact of targeted advertising on advertisers, market access and consumer choice, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2021.

Frau-Meigs D., Societal costs of 'fake news' in the Digital Single Market, European Parliament, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2018.

Koene A., Clifton C., Hatada Y., Webb H., Patel M., Machado C., LaViolette J., Richardson R., and Reisman D., A governance framework for algorithmic accountability and transparency, EPRS, European Parliament, 2019.

Madiega T., Digital markets act, EPRS, European Parliament, 2021.

Madiega T., Digital services act, EPRS, European Parliament, 2021.

Madiega T, Reform of the EU liability regime for online intermediaries: Background on the forthcoming digital services act, EPRS, European Parliament, 2020.

Lomba N., and Evas T., Digital services act, EPRS, European Parliament, 2020.

Marsden C., and Meyer T., Regulating disinformation with artificial intelligence, EPRS, European Parliament, 2019.

Sartor G. Regulating targeted and behavioural advertising in digital services: How to ensure users' informed consent, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, 2021.

Sartor G., The impact of the General Data Protection Regulation (GDPR) on artificial intelligence, EPRS, European Parliament, 2020.

Sartor G., New aspects and challenges in consumer protection: Digital services and artificial intelligence, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020.

Whereas social media provide individuals with new opportunities to access information, express opinions, and participate in democratic processes, they can also undermine democracy by distorting information, promoting false stories and facilitating political manipulation.

This EPRS paper provides an overview of the key risks social media pose to democracy related to surveillance, personalisation, disinformation, moderation and microtargeting. It also discusses key approaches to tackling social media risks to democracy in the context of EU policy.

QA-08-21-390-EN-N

This is a publication of the Members' Research Service
EPRS | European Parliamentary Research Service

PE 698.845
ISBN 978-92-846-8802-9
doi:10.2861/135170