International Journal of Technoethics
Volume 14 • Issue 1

Which Democratic Way to Go?
Using Democracy Theories in Social Media Design

Roxanne van der Puil, Eindhoven University of Technology, The Netherlands
Andreas Spahn, Eindhoven University of Technology, The Netherlands
Lambèr Royakkers, Eindhoven University of Technology, The Netherlands*
https://orcid.org/0000-0001-5161-5855

ABSTRACT

There are concerns amongst researchers and the general public that social media platforms threaten democratic values. Social media corporations and their engineers have responded to these concerns with various design solutions. Though the objective of designing social media democratically sounds straightforward, the concrete reality is not. The authors discuss what a democratic design for social media platforms could look like by exploring two classical conceptions of democracy, one in the liberal tradition and the other in the deliberative tradition. In particular, they discuss three concerns: 1) mis- and disinformation; 2) hate speech; and 3) the relations between filter bubbles, echo chambers, and public debate. By describing the underlying ideals of the two traditions and translating these into design guidelines, the authors make explicit how varied and contrary the implications of different conceptions of democracy can be for addressing public concerns and designing for democratic social media. With these things in mind, this article responds to a call, which is to raise awareness among social media corporations, engineers, and policymakers about varying democratic ideals and the implications that these may have for social media.

KEYWORDS

Autonomy, Democracy Theory, Equality, Ethics of Technology, Liberty, Social Media Design

1. INTRODUCTION

Citizens use social media platforms to be informed, share their viewpoints, and engage with others. Simultaneously, these platforms have come under growing scrutiny and pressure from the public to better regulate the use of social media through design. Three concerns, in particular, are hate speech and bullying on social media, false and misleading information, and the question of whether users should be encouraged to debate with those who hold opposing viewpoints (see, for example, Guiora & Park, 2017, on hate speech; Farkas & Schou, 2019, on post-truth and fake news; and Pariser, 2011, on filter bubbles). In response to these concerns, social media companies have implemented various new design features. For example, users are discouraged from bullying with questions such as 'Are you sure you want to post this?';Meta works with impartial fact checkers who review and rate content

DOI: 10.4018/IJT.331800 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

on Facebook, Instagram, and Whatsapp;[1] users are referred to alternative sources; and Twitter has Community Notes[2] – a community-based approach to addressing misleading information. These design solutions, however, are not uncontroversial. When social media platforms blocked the account of former President Donald Trump in response to the Capitol raid on January 6, 2021, newspaper headlines expressed concern that silencing a President in such a way might be undemocratic.[3] This episode illustrates how complex the notion of democracy is, conveying, for example, both the values of safety and free speech and the range of questions to be answered when designing social media for democracy. Should we design for free speech, or is censoring speech – perhaps even users – at times justified and democratic? Should platforms ensure that all users have an equal voice online through design mechanisms? What functionality should we design recommendation algorithms for? In short, what design choices should engineers make if they are to build a (more) democratic social media?

The public concerns about social media platforms and related design choices are discussed at length in the academic literature.[4] However, to our knowledge, there is little literature that discusses social media design principles as a whole, viewed in light of different theories of democracy. Some notable exceptions are Dahlberg (2011) and Bozdag and van den Hoven (2015). Dahlberg's paper sketches four different democratic theories (liberal-individualist, deliberative, counter-publics, and autonomist Marxist) and their relation to digital technologies, but he does not go so far as to investigate specific design choices for social media. It is exactly this task that Bozdag and van den Hoven (2015) call for as they reflect on the design solutions to the so-called filter bubble effect. They argue that in order to strengthen and diversify designs solutions, engineers should be exposed to various traditions of democracy that embody different democratic norms and thus bear alternative implications for the (re)design of social media. There are a variety of different philosophical approaches to democracy, ranging from branches of classical liberalism (which emphasise the rights and freedoms of individuals) to libertarian accounts (which stress the importance of freedom and opt for minimal governmental intervention) to deliberative approaches (which emphasise the importance of public deliberation over the mere aggregation of votes) to more recent calls for 'radical democracy' (which emphasise the agnostic character of public debate and are sceptical of enlightenment ideals of rationality). Given the scope of this paper, we cannot explore all these different accounts. Rather we choose to focus on two prominent theories. By exploring the traditions of liberal democracy and deliberative democracy, we underline the call by Bozdag and van den Hoven (2015) and illustrate how varied the design directions can be when we design for a democratic social media.

We focus on these two conceptions of democracy for several reasons. the work by philosophers in the liberal tradition, such as Mill and Locke, and the deliberative tradition, such as Habermas, Cohen and Mansbridge, have been of paramount importance for our modern understanding of democracy. While there is also an overlap between the liberal and the deliberative tradition, these views diverge with regard to some aspects of democracy. For instance, whereas Mill's work strongly emphasises freedom, and freedom of speech in particular, Habermas argues for positive communication norms. These two philosophies, applied to social media design, would yield conflicting results. Thus, the works of both traditions help to illustrate the range of design implications that are possible when designers aim for democracy in social media. These theories also partially reflect the current public debate between conservatives who emphasise freedom of speech and progressives who argue for certain limitations (Lakier, 2021).[5] While we recognise and briefly discuss the main criticisms that these traditions have received, we refrain from an in-depth debate on democratic theory. Instead, our aim is to offer a kind of thought experiment whereby we use these theories of democracy to demonstrate the diversity of their implications for designing democratic social media platforms. We expect that future research will continue this thought experiment and explore more theories of democracy. This will help to make designers and regulators (more) aware of the type(s) of democracy to which their design activities contribute. Thinking critically about the design choices and implications of social media design is crucial, as design choices affect online and offline behaviour and attitudes.

We proceed in the following way. First, we discuss both theories, focusing primarily on the work of John Stuart Mill and Jürgen Habermas, and identify their key underlying moral ideals. We then address, one by one, hate speech, misinformation and disinformation, and the question of whether or not to encourage debate. For each of these topics, we first deduct design guidelines from both traditions of democracy. Second, we engage with the ongoing efforts discussed within the behavioural and design literature and illustrate which design choices fit these guidelines. We end each topic section with a discussion, flagging questions for further inquiry. The goal of the paper is not to take a stand on this issue, but to illustrate the implications of two diverging design directions and thus demonstrate how important it is for technology designers to reflect on the key ideals of different conceptions of democracy. Awareness of this issue will bring clarity to public discussions and to issues of uncertainty for social media corporations and their engineers who are designing for a democratic social media.

2. TWO THEORIES OF DEMOCRACY

In this section we will give a short account of classical liberal democracy and deliberative democracy, mainly based on, respectively, John Stuart Mill and Jürgen Habermas. On the one hand, Mill and Habermas both arguably belong to the liberal tradition of democratic theory, in that they combine a strong commitment to the rights and liberties of the individual with a defence of democracy. On the other hand, deliberative democracy theorists, such as Habermas and Rawls, add a strong emphasis on deliberation as the weighing of interests, values and concerns of the public, based on a contractualist political philosophy.

Philosopher John Stuart Mill put forward a utilitarian moral theory, and in the book *On Liberty* he explains and defends the 'no harm principle' (Mill [1859], 1991). The contemporary philosopher and sociologist Jürgen Habermas addresses the public sphere and what he coins 'communicative rationality'. According to Mill, the main principle to defend in democracies is the right to individual and group liberty, with minimal state interference. Conversely, Habermas argues that we should strive for communication norms that help society achieve rational consensus (Habermas, 1996). While Mill's work presents one negative goal, namely what communication should not consist of, Habermas's work defends positive goals, namely what communication should consist of.

2.1 Classical Liberal Democracy

Many of the fundamental principles of liberal democracy can be found in the work of John Stuart Mill. Mill ([1851], 1991) defended individual and group liberty to think, speak and be as one desires and with minimal state interference. He presented several arguments to defend this position. Mill argued that individuals needed extensive liberty to develop their own ideas and identity (Table 1, IB). Only if individuals can develop their own individuality can they pursue their own happiness and ways of living (Table 1, IA). Individuals should be free to believe or not believe in a God; be free to associate with some and distance themselves from others; and live life as they desire. Without the liberty to develop oneself and pursue individual happiness, energy and creativity are lost at the societal level too, to the detriment of societal progress. Moreover, individuals should have the liberty to unite. Finally, the freedom to think and speak freely is important for the discovery of truth (Table 1, IB). Mill was very much concerned with individuals' cognitive skills and knowledge acquisition. He explained that the assertion of both true and false claims is beneficial when it comes to successfully distinguishing between well-grounded arguments and charlatanism, and maintaining meaning in truth. It is important to distinguish false from true claims and to revise true claims based on new insights. Thus, both true and false claims stimulate debate and aid the development of cognitive skills.

It is, however, important to note that for Mill, individual liberty offers protection not only against the state, but also against majority views (Table 1, IC). Mill warned that a majority could be as controlling and tyrannising as a sovereign. Without the protection of individual liberty, there

Table 1. A paradigmatic overview of the key ideas in two strands of democracy theory

	Liberal Democracy	Deliberative Democracy
A. Main principle(s)	Freedom (from interference) to choose your own (private) version of the good life	Communicative rationality: rational agreement, truth, autonomy and equality
B. Underlying motivation	Protecting individuality and self-development, pursuit of truth	Establishing mutual understanding and rational agreement on the best way to organise society
C. Main threats	Illegitimate interference by the government, private actors such as corporations, and popular opinion (tyranny of the majority)	Illegitimate influences on a rational, free and equal debate (coercion, misusing one's power to dominate, false information, etc.)
D. Implications for social media companies	The only restriction on companies' design of social media platforms is that they must remove content that incites violence and content that leads to substantial harm, including mental harm	Companies should strive towards designing social media platforms in such a way that an ideal speech situation is encouraged (i.e., users engage in a truthful, rational, free and equal discussion)
E. Implications for social media users	Users are free to interact on social media and 'naturally punish' other users as they please, provided they do not incite violence or cause substantial mental harm	When interacting with others online, users should intend to be truthful, rational and respectful of others to reach understanding

is a threat of social conformism induced by a majority effect. For these reasons Mill concluded that there is only one reason to interfere with the freedom of individuals, which is summarised in the ´no-harm principle´. This principle asserts that government interference in the life of individuals and groups is legitimate only when it prevents direct, physical harm to others. '[T]he only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others. His own good […] is not a sufficient warrant' (Mill, 1991, p.30). Thus, generally speaking, Mill was not in favour of paternalism.[6] At the time that Mill wrote his work, mental harm induced by speech was not yet recognised and therefore offered no justification for governmental interference. Mill did, however, promote the idea of natural punishment. People, rather than the state, should correct each other when individuals or groups think and say things that others consider wrong, foolish or immoral. Forms of natural punishment include expressing an opposing opinion, social stigma, public contempt, avoiding a person or warning others to avoid him. These forms of correction should not be paraded publicly, Mill argued, but can help to promote truth and a morally better society.

In sum, the fundamental ideal of democracy according to Mill is freedom of individuals and groups to think, say and act as they desire as long as this does not incite physical violence, thereby harming others (Mill, 1991). Whereas natural punishment by individuals or groups, such as companies, should be part of a functioning democracy, legal punishment or censorship by the state when speech does not inflict direct physical harm is undemocratic. This freedom helps prevent social conformism and promotes individuality, happiness, societal progress and truth. Later developments of liberal democracy theory have emphasised the importance of liberty and equality for true democracy (Hösle, 2004, p. 639ff; Doomen, 2014) and defended liberal democracy as a fundamentally superior system of governance (Fukuyama, 2006).

But various criticisms have been brought forth against these theories as well. The public conversations discussed by Mill took place in public spaces. Times have changed. Even though traditional elements of the public sphere still exist outside of the internet, the new digital platforms on which people debate are controlled and owned by social media corporations (Habermas, 2022). Modern liberal theories must take these developments into account. Moreover, we now know that

harm can be both physical and mental. Philosophers have updated Mill's work to include these facts. However, the ways people can disagree over harms are prolific: should insults to prophets whose sacredness is not generally recognised – or forms of art that some endorse but others see as expressions of decadence or immorality – be considered as harm (Valkenburg, 2013)? Bell (2021), for example, discusses what speech deserves restriction if a broader notion of harm is applied. According to Bell, bigoted speech and speech that promotes 'the love of domineering over others' – which Mill considered as an 'immoral disposition that justifies moral disapprobation' (p. 179) – can, in certain situations, induce mental harm that should be restricted. In this article we account for the recognition of mental harm. We will define mental harm as substantial harm to a person's psychological or intellectual functioning; such harm may be evidenced by a substantial degree of characteristics such as anxiety, depression, withdrawal or outwardly aggressive behaviour (see, e.g., Amoretti & Lalumera, 2019). One offensive remark is generally not sufficient to cause substantial mental harm, but many offensive remarks of a similar kind directed at one individual or group do have the potential to cause substantial mental harm.

2.2 Deliberative Democracy

The concept of deliberative democracy is a significant idea in contemporary political philosophy. Its proponents aim to encourage citizens to actively engage in the democratic process by reflecting on important issues. Habermas describes the ideal conditions of public deliberation, also referred to as the ideal speech situation. To this end Habermas distinguishes two notions of rationality: strategic and communicative reason. A rational person is someone who knows how to reach her ends. Strategic rationality lies at the heart of 'manipulation', either of nature or of other humans if they are used – in Kantian terms – as mere means. This form of rationality is rooted anthropologically in the human need to manipulate nature to survive, as in hunting or farming. Communicative reason, on the other hand, is rooted in the fact that we are social animals that need to jointly establish and agree upon the rules by which we want to live (Habermas, 1984). Communicative rationality, then, concerns the questions of which goals we should set for ourselves, either as individuals or as a community. Habermas (1996) argues that in a true democracy, when making societal decisions, we should rely on the normative principles of communicative rationality (Table 1, II.A). When the normative principles of communicative reason are met, we have realised an ideal speech situation, which comprises the following: freedom and equality for all citizens to partake in the deliberation process; freedom and equality for all citizens to question whatever has been stated and to be critical of one another's arguments; the intention of being truthful; having an open attitude towards others when partaking in deliberation; and the avoidance of instrumental reason (Table 1, II.C). Thus, in an ideal speech situation, citizens attempt to reach a mutual understanding or agreement based on equality and freedom and the 'power of the better argument' only as opposed to the power of manipulation (Table I, II.B). Habermas acknowledges that oftentimes public debate is non-ideal and distorted. This is why the availability of information should be encouraged and why legitimate institutions and procedures are necessary. The only legitimate institutions and procedures are those which all citizens would agree on, allowing individuals to be self-determining in the way they are governed. (Habermas, 1996). In sum, the conditions of an ideal speech situation, together with legitimate procedures and institutions, reinforce public deliberation characterised by communicative reason, thereby encouraging mutual understanding and agreement as to how society should be organised. Thus deliberative democracy achieves two goals. First, it encourages democratic virtues such as societal cooperation, autonomy, rational deliberation, mutual respect and equality by striving towards the ideal speech situation. Second, it allows agreement or mutual understanding as to how society should be organised.

The later tradition of deliberative democracy theories has upheld the importance of key ideas, such as the importance of participation, equality, the absence of power asymmetries, the ideal of mutual respect, and of course, the central role of deliberation for democratic decision-making processes

(Bächtiger et al., 2018). However, other ideals have come under increasing scrutiny or have been modified, in particular the idea of consensus. This ideal has been criticised for being too demanding (Rienstra & Hook, 2006). Deliberative democracy theory has been criticised for ignoring the reality of power dynamics and the way in which people are situated in different positions of power within society (Mouffe, 1999; 2005). According to the Mouffian critique, Habermas's view of communicative action fails to take into account the underlying power dynamics that exist in society and therefore fails to provide an adequate account of how communication actually works in practice.

Social media in particular is far removed from an ideal public sphere, and obtaining such an ideal in this context may simply be impossible. Crawford (2016) has taken up the Mouffian critique of Habermas in the digital sphere. Crawford argues that Habermas's concept of the public sphere fails to consider how digital media can be used to manipulate public opinion and shape discourse. Digital media, such as social media and search engines, can be used to curate and control public discourse in order to benefit particular individuals and groups, thus undermining the concept of a public sphere. The Digital Service Act,[7] which comes into effect in February 2024, is a significant tool to regulate the digital space and grapple with the issues of power asymmetries and users' rights. Moreover, rational consensus has been criticised for clashing with the reality of widespread voter ignorance and irrationality (Somin, 2010). Rational irrationality, i.e., a situation where it is instrumentally rational to lower epistemic standards because of a strong attachment to one's views and preferences (Caplan, 2001), reduces the quality of decisions made through deliberative processes, because individuals may prioritise other factors over truth-seeking. For instance, individuals could adopt a bad policy because the arguments for it align with their pre-existing prejudices or are emotionally satisfying (Caplan, 2001).

In reaction to this criticism, proponents of deliberative democracy theories have emphasised consensus as only one possible goal of deliberation, along with less ambitious goals, such as clarifying conflict or reaching a fair compromise (Mansbridge et al., 2010; Habermas, 2022), while acknowledging the importance of disagreement (Gutmann & Thompson, 2004). Others have pointed out that the principles of deliberative democracy are meant to be understood as ideals. Under this conception, either we should strengthen these ideals in social media design, despite the influence of power and irrational behaviour (Fuchs 2015), or we should distinguish different models of epistemic deliberation within a democracy (Marrone, 2022). Similarly, Dahlberg (2001) has applied a deliberative model in the spirit of Habermas to investigate the prospects of online deliberation.

2.3 Summary: Key Principles of Liberal and Deliberative Democracy

While we keep these criticisms in mind, we use the original work of Mill and Habermas as the main inspiration for our discussion of social media. However, when discussing the design implications of these theories in the following section, we do raise questions for further inquiry pertaining to these criticisms. Moreover, we encourage others to consider recent work that responds to these and other criticisms directed at both the liberal and deliberative views of democracy. The differences between Mill's and Habermas's theories of democracy will have a significant impact on the design requirements of social media platforms. The emphasis on individual liberty in liberal democracy theories will mean that social media platforms should be designed to protect users' rights and to limit governmental control (as well as control from other actors). On the other hand, the emphasis on open dialogue and communication in deliberative democracy theories will mean that platforms should be designed to facilitate meaningful dialogue between users and to promote mutual understanding and respect. However, both traditions emphasise the importance of users' being able to express their opinions freely, without fear of censorship or retribution, as well as being able to assess the accuracy of content.

3. HATE SPEECH AND BULLYING

One concern that social media platforms try to tackle is that of hate speech and bullying. Social networks and the internet seem to amplify cyberhate, and unfortunately, online hate speech also affects the offline world (Castaño-Pulgarín et al., 2021). Cyberhate results in 'hate crimes, offline aggressions, discrimination, racist attitudes, democratic consequences, exacerbation of gendered violence, among others, which affect coexistence and mental health of victims, bystanders or perpetrators' (p. 5). What is a democratic way to deal with hate speech and bullying on social media platforms?

3.1 Design Guidelines

Both traditions of democracy theory would urge platforms to tackle hate speech—each, however, to a different extent. Liberal democracy emphasises the importance of free speech as a fundamental right, even if it includes offensive remarks and hate speech. It strives to balance protecting free expression while limiting harm by establishing clear rules and guidelines to govern hate speech without impeding free expression. On the other hand, deliberative democracy prioritises fostering constructive and inclusive dialogue. It places a greater emphasis on minimising harm and creating an environment conducive to reasoned deliberation, potentially allowing for more restrictions on hate speech. It emphasises the role of platform governance in actively facilitating and promoting constructive deliberation; in order to ensure inclusivity, it involves users in decision-making processes and platform governance.

The liberal view offers the following design guideline for social media platforms, focusing primarily on protecting individual users from harm: ***L1)*** *only social media speech that directly leads to substantial mental harm or physical harm (i.e., incitement to violence) should be removed.* As Mill's work emphasises liberty, this principle generally allows users to offend one another now and again. However, a trickier prospect is applying this guideline to users who have united on the basis of a common viewpoint or belief. What if people unite against feminism? What if people unite against homosexuality? What if people unite as white supremacists? This guideline could be interpreted as follows: as long as such expressions are kept private – within the user group – no substantial mental harm can ensue, and therefore, corporations should not interfere. In practice, however, such beliefs and viewpoints are rarely kept private, and behaviour is shaped by one's beliefs and opinions. Thus, according to the liberal view, social media platforms are not obligated to remove speech that many people may find offensive unless it can be argued that it causes direct harm.[8]

Contrary to this view, a deliberative view urges platforms to go further and places a greater emphasis on the role of social media platforms in promoting democratic values, facilitating public discourse, and giving users equal access to the platform and an equal opportunity to participate in public debate. Platforms should therefore not approve of groups that express exclusionary attitudes towards some users, even if these groups do not act on these attitudes. These groups and such content disrespect the principle of equality. Thus, a deliberative view on democracy offers this design guideline: ***D1)*** *users should respect the principle of equality when interacting online.*[9] Hence, content moderators directed by a deliberative perspective would be strongly encouraged to remove, or at least make less visible, all statements such as 'Kill president X!' and 'President X is dumb, fat and ugly' and to seek to prevent users from uniting on exclusionary grounds. Content moderators directed by a classical liberal perspective, on the other hand, would remove statements of the first kind but would accept statements of the second if they are isolated and not repeated. They would interfere with groups only if they express their exclusionary attitudes towards others in directly harmful ways. Similar arguments for content moderation could be made in the case of coercive or deceptive content.

3.2 Design Options

To remove content inciting violence, which both traditions would direct moderators to do, engineers could develop and integrate cyber-aggression detection algorithms into social media platforms.[10]

However, this popular solution faces two problems. First, to prevent violence, violence-inciting content must be detected *before* it is shared via social networks. Second, nuanced interpretations of content by algorithms are rare, and the detection of content inciting violence is difficult. Algorithms struggle to distinguish between factual claims and opinions, between humour, sarcasm and irony, and between extremist content and counter-extremist content (Lorenz-Spreen et al., 2020). Some statements made by Donald Trump on January 6 illustrate these problems. Does the statement 'And we fight. We fight like hell. And if you don't fight like hell, you're not going to have a country anymore'[11] qualify as inciting violence? This question was debated only *after* Donald Trump spoke to millions of people. Friction methods, i.e., delaying the time between writing and publishing content, can help address these issues while preserving freedom of speech. One friction method is *quarantining speech* (Ullmann & Tomalin, 2020), in which algorithms first detect content that may incite violence and put it on hold, giving moderators the time to review content carefully. This allows moderators to pre-emptively remove content and thus prevent physical harm. Engineers can also integrate *nudges* to prevent incitement to violence (Thaler & Sunstein, 2008). Nudges are 'interventions designed to steer people in a particular direction while preserving their freedom of choice' (Hertwig & Yanoff, 2017, p. 973). If algorithms and/or moderators predict that a message in the making may incite violence, pop-up messages can inform and nudge the user to change the content. An example of this is OpenWeb, with Jigsaw's Perspective API.[12] Each comment submitted by a user is subject to moderation through an algorithm. If the content is deemed problematic or if it does not comply with the publisher's Community Guidelines, the system will not immediately reject it or require additional human review. Instead, the user will receive a message encouraging them to review and reconsider their comment. The user has the option to edit their comment and resubmit it or to post it anyway and accept the consequences. The system allows users to review and edit their comments before submission in order to prevent them from attempting to manipulate the system. Each comment is granted a single opportunity for revision. Research suggests that this method encourages a healthier conversation, whilst minimising censorship (Simon, n.d.).

To respond to the liberal guidelines for removing content that causes substantial mental harm, social media corporations could use detection systems that track down individuals and groups who offend, bully and make racist remarks not once, but on many occasions, to the same individual or same group of individuals. Another way to mitigate physically and mentally harmful content from entering social networks is to broaden the range of available emotional expressions online, such as, e.g., the use of emojis. While paralinguistic and emotional cues are important for communication, their availability is limited online. One empirical study found that people are more overconfident in evaluating the emotion of a message when it is transmitted by email than when it is transmitted by voice or face-to-face communication (Kruger et al., 2005). The findings of this study suggest that it is not the gesture and expression of people that accounts for the difference in confidence, but rather the lack of intonation and vocalisation. Another study (Kraut et al., 2009) found that media which supports speech is more likely to prevent emotional escalations than media which supports only text. In light of this and other research on emotions, Marin and Roeser (2020) argue that media should be designed for richness and complexity of emotions so as to aid nuance and self-expression and contribute to sympathy and understanding of shared values. They are hopeful that redesigning platforms in such a way will make participants more aware of the emotional consequences their actions may have and that 'online debates could become more meaningful for all participants, possibly leading to a form of digital civic well-being (p. 148). One way to accomplish this would be to allow users to post or respond with video and audio clips rather than short text-based messages. Design options that offer greater emotional nuance with the aim of keeping discussions civil and increasing sympathy should be tested and adjusted, to prevent the reverse effect from materialising.

As the deliberative view promotes positive ideals for communication, guideline D1 speaks to the intention of users too. Users should not merely respect the principle of equality; they should intend to respect equality. To strengthen the democratic design, social media platforms could run online

campaigns for nonviolent communication (NVC). NVC promotes communication that is respectful of equality and liberty (Rosenberg & Chopra, 2015). Moreover, as the intention of users matters, algorithms and moderators should weigh this into their review process. Trolls, for example, intend to offend or confuse other users. As DiFranco (2020) puts it, the act of trolling is pro tanto wrong, as its intention is to disrespect the principle of equality. As such, Guideline D1 directs platforms not only to remove acts of trolling but also to detect those who intend to troll, with the goal of changing their intentions and behaviour. In recent years there has been much effort to improve the detection of trolls (see, e.g., MacDermott et al., 2022).

3.3 Discussion and Further Research Questions

While the guidelines for removing incitement to violence and content that is substantially harmful on a mental level seem straightforward in theory, they need to be elaborated further. In order to remove and discourage incitement to violence, engineers first need to determine exactly what counts as such. As scholars such as Rauch (2021) point out, cancel culture has established itself firmly in societies and poses a severe threat to liberal democracy. Free speech, which includes the ability to say things which may offend others, is crucial for any liberal democracy. To address both hate speech on the one hand and cancel culture on the other, several questions need to be examined further. When do words cause harm? How often does something need to be said in order for it to cause harm? When must offensive and harmful comments be permitted in order to protect free speech, and when should such comments be regulated? The same questions are raised when platforms try to remove and discourage content that intends to disrespect and is harmful in a broader sense. Though the intention matters to Habermas, determining the intention of users seems rather difficult, if not, at times, impossible. It may be easy to detect bots with bad intentions, but correctly detecting the intention of communication in the case of irony, humour or harmless trolling might be more difficult. And what about users who express their emotions without careful reflection, who do not intend to harm others? How should such content be judged?

4. MISINFORMATION AND DISINFORMATION

Another societal concern is the spread of misinformation and (political) disinformation via social networks. Misinformation is misleading information that is shared online *without* the intention to mislead, and disinformation is false information that is shared online *with* the intention to mislead (Skyrms, 2010). Such information presents a challenge for democracies, as democracies represent the ideal of self-government by the people. In order for people to self-govern, they need to be informed correctly, or at least have the opportunity to receive accurate information, and be free from manipulation and coercion when making decisions. Only if these conditions are met can people make autonomous decisions and truly be capable of self-government. What is a democratic way for engineers to respond to the problem of mis- and disinformation on social media platforms?

4.1 Design Guidelines

The strategies that the work of Mill and Habermas promote are partly similar, and partly distinct. As, for Mill, the no-harm principle dictates the boundaries of legitimate state interference, moderators guided by this conception of democracy would leave false and misleading information unregulated. Mill´s reasoning for this principle, as explained in Subsection 2.1, was not that truth is not worthy of protection, but rather that free speech promotes truth and safeguards democracies. One could thus formulate the following design guideline: ***L2)*** *false information should not be removed from social media platforms.* In addition, as Mill did stress the importance of knowledge and the development of cognitive skills, there would be mechanisms in place that make content assessable by users of the platform. Since Mill's defense of free speech is aimed at allowing users to freely engage with each other and correct their biases and false beliefs, an additional guideline could be taken on board

by designers, namely ***L3)*** *Social media platforms are encouraged to promote intellectual skills through design.* Strengthening these skills can help users identify what information is credible and what information is not. Moreover, Mill encouraged citizens to respond correctively to comments and behaviour of other users if these are immoral or foolish, and false and misleading information qualifies as such. This directive can be translated into the design guideline ***L4)*** *to address foolish or morally problematic statements, users should be enabled to correct one other.* While L2 is strictly required by Mill if social media corporations are to design a democratic platform, L3 and L4 would be strongly encouraged.

The ideal promoted for society, or in this case, for the online public, by Habermas, is for citizens to reach an agreement based on the best argument alone. To achieve this goal, strategic and manipulative reasons ought to be avoided and users should intend to be truthful. Another way to phrase this is that online communication should respect the principle of individual autonomy. Though the concept of autonomy has been debated for centuries, we use a general definition of the word. People are autonomous if their choices are not interfered with and meaningful decision-making is not constrained, for example, through inadequate understanding (Beauchamp & Childress, 2009). To achieve and encourage online communication that respects autonomy, two guidelines should be considered. The guideline ***D2)*** *misinformation and disinformation should be removed from social media platforms* ensures that users are well informed and not deceived or manipulated by external influences. The guideline ***D3)*** *Social media platforms are encouraged to promote intellectual skills through design* further promotes autonomy in the online public sphere. Triggering and developing the cognitive skills of users can help users to distinguish true content from false and manipulative content. Hence, while both Mill and Habermas agree that engineers should promote the use and development of users' cognitive skills through design, they disagree on whether or not mis- and disinformation should be removed. However, both traditions value truth and rationality; therefore it is important for both of them that assessability mechanisms be established.

4.2 Design Options

We will start with design options that fit the design guideline both traditions agree on: L3/D3) *Social media platforms are encouraged to promote intellectual skills through design.* Designers can rely on nudging and boosting techniques to do so. Nudges require no effort on the part of the user and, as explained in Subsection 3.2, steer people's decisions while preserving freedom of choice. Boosting interventions aim to optimise people's cognitive and motivational abilities and, unlike nudges, do require some effort on the part of the user (Lorenz-Spreen et al., 2020). The idea behind boosting interventions is that cognitive competences are internalised over time, so that eventually, interventions are no longer needed (Hertwig & Grüne-Yanoff, 2017). Nudging and boosting techniques preserve freedom of choice while stimulating critical thinking, thereby fitting the underlying ideals promoted by the liberal view. They fit the ideals of the deliberative view by encouraging users to think carefully about the information they encounter online, thus promoting autonomy.

A nudging technique that fits guideline L3/D3 is to display more information metrics. When users are only presented with like and dislike statistics, users can, for example, mistake the number of likes for a majority opinion or societal consensus. This is also referred to as false consensus, i.e., when one sees one's 'own behavioral choices and judgments as relatively common and appropriate to existing circumstances while viewing alternative responses as uncommon, deviant, or inappropriate' (Ross, Greene & House, 1976, p. 280). False consensus can reinforce user opinions (Lorenz-Spreen et al., 2020). If information such as the total number of users who have scrolled over a post and the total amount of time users spend reading an article appears on users' dashboards, for example, users are nudged to think more carefully about what this information suggests about the public's opinion. As people may be nudged to opt for healthier products in supermarkets by the placement of fruit rather than sweets at the counter (Thaler & Sunstein, 2003), social media users can be nudged to

consult credible sources and information centres by placing referrals to such sources strategically or including a label or signal about the reputability or the political leaning of a source.

A boosting technique that fits guideline L3/D3 is to enable users to decide how their newsfeed is organised and designed, making them the architects of their personal social media environment (Lorenz-Spreen et al., 2020). As Pennycook et al. (2021) point out, oftentimes, users are not encouraged to reflect on the accuracy of posts. Therefore, a second boosting intervention is to teach users the skill of lateral reading. Professional fact checkers use this skill. In lateral reading, other websites and resources are used to critically assess the credibility of information presented in one source. When lateral reading is boosted through design, users come to internalise questions such as 'Who runs this site?' and 'What evidence supports this claim?' To boost lateral reading, engineers can integrate pop-up questions and decision-trees into social media platforms. When the user has answered the question(s), a final pop-up message shows the user whether the source is likely to be reliable or not based on the answers provided by the user (Lorenz-Spreen et al., 2020). Third, designers can periodically ask users to evaluate the accuracy of a random selection of headlines (Pennycook et al., 2021). If users are asked to explain why they believe a headline is true or false, they become less inclined to share those with false headlines (Fazio, 2020).

Although both the classical liberal and deliberative view are committed to rational truth seeking, their methods, as expressed in guidelines L2 and D2, might be in tension. Whereas the deliberative view would strongly encourage designers to remove false and misleading information, the classical liberal view tends to warn designers of the dangers of limiting free speech. Social media corporations directed by the deliberative view might employ third-party fact checkers and algorithms to detect and remove mis- and disinformation spread by users and bots, to make for a more democratic social media, whereas platforms directed by the classical liberal view might be tempted to give more weight to freedom of expression. As expressed in L4, social media engineers directed by a liberal view would encourage users to speak up when others share mis- and disinformation. Research has found that the likelihood of users' correcting other users is affected by different factors, including 'one's relationship with the user who posted the fake news article' (Tandoc, Lim & Ling, 2020, p. 389). Users are more likely to respond to fake news posted by friends and family members. Thus, to design for L4 and address mis- and disinformation, engineers should prioritise posts from friends and family members on users' newsfeed pages.

4.3 Discussion and Further Research Questions

The empirical evidence pertaining to removing and flagging misinformation offers arguments to support designing both in the direction of Mill and in the direction of Habermas. While using algorithms and fact checkers to review content can be successful, oftentimes mistakes are made. As explained earlier, algorithms – and, at times, people – struggle to accurately distinguish between factual claims and opinions and to identify the intent behind a post – for example, whether a post is meant sarcastically or seriously. Moreover, at times detailed or expert knowledge is needed to review content accurately. Niemiec (2020) discusses moderation of COVID-19-related content and explains that in this context specifically, critical questions are pertinent to minimise or remove the risk of drawing the wrong conclusions from data. Yet, as she shows, there are plenty of instances when videos of researchers raising legitimate critical questions – for example about the lockdown – have been removed. While there is a desire to prevent conspiracy theories from spreading through social media platforms, at times it is difficult to draw a clear-cut boundary between conspiracy thinking on the one hand and healthy skepticism and rational critique of science on the other (Huneman & Vorms, 2018). This means that when content moderators and algorithms get it right and successfully remove false and misleading content, truth *is* promoted and autonomy *is* respected, in the sense that people cannot be misinformed by that item of content. Yet when content moderators and algorithms get it wrong, the process of finding the truth and autonomy is undermined. While one can argue from a consequentialist standpoint that moderators and algorithms get it right more often than not and thus

that removing such content leads to a better outcome, this standpoint is morally controversial and would benefit from further debate. Moreover, empirical research suggests that flagging content as (potentially) false or misleading does not automatically trigger users to think critically (Gaozhao, 2021, p. 10). People tend to accept flags and do not tend to investigate the accuracy claim, even if flags are misplaced. In addition, if posts are not flagged and the content does not contradict existing beliefs, users consider those posts to be true even if they have simply not been flagged yet (Gaozhao, 2021). These findings raise questions about what is the morally best and most democratic strategy for social media engineers to use in combating mis- and disinformation.

5. FILTER BUBBLES, ECHO CHAMBERS AND ENCOURAGING PUBLIC DEBATE

A third topic of concern is whether engineers should encourage debate between users, and if so, how. Many researchers have investigated the effects of so-called filter bubbles and echo chambers (Pariser, 2011). Filter bubbles have the supposed effect of keeping people in an information bubble that is one-sided rather than multi-sided.[13] An echo chamber is the online situation in which certain beliefs are repeatedly reinforced and amplified, rather than challenged. The concern here is that instead of receiving a wide variety of opinions and perspectives, users encounter only opinions and perspectives largely aligned with their own viewpoint, and that that this imbalance will harm democracy and exacerbate polarisation. In response, many design solutions have been developed to counter the effect of filter bubbles and echo chambers and encourage interaction between users with opposing viewpoints. If social media platforms are to be democratic, who should engage in debate with whom online, and how should users engage in debate with one another?

5.1 Design Guidelines

On this topic too the traditions offer social media corporations different design directions. Classical liberal democracy recognises the importance of individual freedoms, including freedom of expression and freedom of choice. From a classical liberal democratic view, filter bubbles may not be inherently problematic, as they arise from individuals' autonomous choices to seek out and consume content that aligns with their preferences. Classical liberal democracies tend to prioritise the protection of individual rights, including the right to access and assess information and express oneself, even if it leads to the formation of filter bubbles. They may focus on providing individuals with tools and resources to navigate and customise their own information environments. But from a deliberative democratic view, filter bubbles are seen as problematic as they hinder the exchange of diverse perspectives, limit opportunities for informed deliberation, and contribute to social polarisation. Deliberative democrats argue that filter bubbles can undermine the deliberative process by narrowing the range of views and by inhibiting the formation of shared understandings. They advocate for interventions that actively counteract filter bubbles, such as algorithmic transparency, diverse content exposure, and public deliberation spaces.

As Mill argued that people should have the liberty to unite with other people for whatever purpose they see fit, as long as this does not lead to physical harm or – to adjust Mill's theory to current knowledge – substantial mental harm. Filter bubbles and echo chambers are problematic only if in those bubbles and chambers, speech is used that leads to substantial mental or physical harm. This means that guideline ***L1)*** *only social media speech that directly leads to mental harm or physical harm (i.e., incitement to violence) should be removed*, is applicable to this topic as well. Even if other users dislike the grounds of group formation – for example, a disgust towards working women – social media companies are not obligated to discourage this group, unless substantial mental harm is inflicted, nor to expose them to people who think differently. However, the guideline promoting natural punishment is relevant here again too: ***L4)*** *to address foolish or morally problematic statements and online groups, users should be enabled to correct one another.* If users feel individual or group

attitudes – e.g., the belief that women should not work – are morally problematic, those users should be encouraged to speak up.

Conversely, deliberative democratic theory would direct engineers to encourage interaction between those holding opposing views, to promote mutual understanding and consensus. Moreover, this interaction should be critical, yet respectful and open-minded. Hence, engineers would design for the following guideline: ***D4)*** *users should be encouraged to deliberate about (alternative) viewpoints with each other whilst having a critical, yet open mindset*. Guideline D1 is relevant to this topic too: ***D1)*** *users should respect the principle of equality when interacting with others online.* Both of these guidelines increase the chances of users' coming to a mutual understanding, as the former aims at bringing users with different viewpoints together and the latter aims at making the interaction civil.

5.2 Design Solutions

To address this topic, engineers directed by deliberative democracy would, through design choices, encourage users to interact with users who hold different viewpoints, whereas engineers directed by classical liberal democracy would leave this choice to users. As discussed previously in Subsection 3.2, there are various design options that fit guidelines L1 and D1, such as using detection algorithms, moderators and nudging techniques. L4 again encourages designers to strengthen user-to-user correction. Designers should design for commenting, as well as providing the option to unfriend and (publicly) block fellow users. Elder (2020) explains the importance of designing for 'defriending'. By defriending, users send out a strong message that they disagree with another user's beliefs or behaviour, i.e., they apply natural punishment, and defriending may even help prevent messaging from becoming extreme. In 2020 Twitter made it possible for users to decide who can and cannot reply to their tweets and join in on a conversation. With this feature Twitter is attempting to further enhance user control.[14] This allows users to naturally punish users who have previously misbehaved,

Table 2. An overview of classical liberal theory- and deliberative democracy theory- informed design guidelines for social media

Classical liberal democracy theory	Applicable to	Deliberative democracy theory	Applicable to
L1: Only social media speech that directly leads to substantial mental harm or physical harm should be removed (*required*)	Hate speech and bullying; filter bubbles, echo chambers and encouraging public debate	**D1:** Users should respect the principle of equality when interacting online (*required*)	Hate speech and bullying; filter bubbles, echo chambers and encouraging public debate
L2: False information should not be removed from social media platforms (*required*)	Misinformation and disinformation	**D2:** Misinformation and disinformation should be removed from social media platforms (*strongly encouraged*)	Misinformation and disinformation
L3: Social media platforms are encouraged to promote intellectual skills through design (*strongly encouraged*)	Misinformation and disinformation	**D3:** Social media platforms are encouraged to promote intellectual skills through design (*encouraged*)	Misinformation and disinformation
L4: To address foolish or morally problematic statements, users should be enabled to correct one other (*strongly encouraged*)	Hate speech and bullying; Misinformation and disinformation; filter bubbles, echo chambers and encouraging public debate	**D4:** Users should be encouraged to deliberate about (alternative) viewpoints with each other whilst having a critical, yet open mindset (*strongly encouraged)*	Filter bubbles, echo chambers and encouraging public debate

i.e., responded in an immoral way to their conversation. Natural punishment or social control online is being studied extensively; see, for example, Hillman et al. (2021). Mill would encourage this research and its application to the (re)design of platforms.

Guideline D4 also encourages users to be critical of one another's standpoints through commenting; yet since it aims at encouraging interaction, it directs engineers not to design for blocking or disabling users from commenting on one's content. Additionally, there are various tools to address filter bubbles and promote the goal of D4. The interface 'Reflect', for example, changes "online comment boards: to the right of every comment, visitors are invited to restate points they hear the commenters making" (Kriplean et al., 2012). This encourages users to have a more open mindset by modifying 'the comments of webpages in order to encourage listening and perspective taking' (Bozdag & van den Hoven, 2015, p. 258). After reading or listening to the comments, users are encouraged to reword what they have just listened to. Furthermore, research shows the importance of word choice for affecting the likelihood of users' reading alternative viewpoints. Yom-Tov, Dumais, and Guo (2014) show 'that when the language model of a document is closer to an individual's language model, it has a higher chance of being read despite it describing an opposite viewpoint' (p. 152). Thus, in engineering recommendation algorithms for guideline D4, this factor should be taken into account. Finally, research shows that creating anonymous, one-on-one conversations between people with opposing viewpoints benefits understanding and increases the chances of people's finding common ground (Dahlberg, 2001; Bail, 2022). Other design options for discouraging filter bubbles are algorithms that intentionally expose users to content representing diverse viewpoints, even if they differ from their own, and content recommendations that consider a wide range of perspectives, including those that challenge users' existing beliefs. An example is FlipFeed[15], a Google Chrome extension, developed at the MIT Media Lab, that empowers Twitter users to substitute the feed of another actual Twitter user for their own feed. This extension utilises deep learning and social network analysis to select feeds on the basis of deduced political ideology ('left' or 'right') and presents them to users. Subsequently, the user has the option to revert back to their original feed or repeat the process with another feed.

5.3 Discussion and Further Research Questions

Empirical literature points out that the effects of filter bubbles may be overstated (see, e.g., Zuiderveen Borgesius et al., 2016). Whether or not this is true, there is also conflicting evidence regarding the effects of tackling filter bubbles and echo chambers and trying to bring people with different viewpoints into closer contact. Bail (2022) argued that the push to break filter bubbles and echo chambers is dangerous and that attempting to do so can exacerbate polarisation if not carried out correctly. He finds that exposing users to content that does not fit their (political) ideals can motivate those users to defend their beliefs even more strongly, while anonymous conversations between two people aid understanding and compromise. The empirical findings on encouraging interaction between those who hold opposing viewpoints show that while engineers may intend to design for certain values, those design choices may actually deliver different results. While exposing users to different viewpoints embodies values such as diversity and openness, if the response of those same users is to defend their beliefs even more strongly, the opposite effect may be realised. In other words, while theories of democracy can convincingly argue which values are important to embed in technologies, a different question is whether or not the design of technologies delivers the desired values (van de Poel, 2021). Whether values are realised depends not only on technical aspects such as design, but also on how users respond to those design features (van de Poel, 2021). The underlying principle of Mill's work is that people can decide themselves how they want to lead their lives, without illegitimate interference. At first glance, therefore, it seems that people should have the freedom to decide whether they want to immerse themselves in filter bubbles and echo chambers or not. However, one can question when this choice is autonomous and when it is technologically induced. Social media platforms are known for their emotional pull and their ability to feed users content that is hard to ignore. For example,

research by Alfano et al. (2021) confirmed the hypothesis that the recommender system of YouTube can promote more extremist and radical content depending on the initial topic that brings people to the platform. It is also known that a great deal of time and money is invested in social media strategies during election periods. Research has confirmed that these strategies can influence public opinion and thereby, the results of elections (e.g., Gorodnichenko et al., 2021). Thus this design principle needs further investigation concerning the long-term effect that social media may have on people's thoughts.

6. CONCLUSION

In this article we addressed various concerns about social media design in relation to democracy, namely hate speech, mis- and disinformation, filter bubbles, echo chambers and stimulating public debate. More specifically, by exploring deliberative and classical liberal democracy, mainly through the work of Jürgen Habermas and John Stuart Mill, we illustrated that differing conceptions of democracy lead social media corporations and their engineers to address these issues with sometimes diverging design strategies, resulting in social media platforms that often differ significantly from one another. This shows how important it is for social media corporations and their engineers, as well as those in charge of policymaking, to become (more) aware of the democratic norms that they are designing for. Admittedly, an awareness of these theories alone will not solve these issues, since there are many complicating factors at play, such as economic pressures on traditional and new media platforms. An in-depth analysis of these factors is beyond the scope of this paper. However, an awareness of the ethical implications and the underlying theory is a first step. This allows social media companies to consciously position themselves in terms of how they view democracy and to subsequently implement rules which fit this position. At the same time, we showed that while the insights of classical theories of democracy are fruitful in starting the discussion on democratic social media design and in flagging questions for further research, they do not offer all the answers. While the design guidelines offered by the traditions are largely divergent, some similarities can be found too. Both theories require engineers to remove speech acts inciting violence. Yet to truly design for a democratic social media, the deliberative view tells engineers to go further by discouraging and removing all forms of harmful content and by encouraging interaction that respects autonomy and equality. This means that online groups that do not respect the principles of autonomy and equality would be removed. When it comes to mis- and disinformation, both traditions present guidelines to strengthen the cognitive skills of users. Both positions would therefore defend design choices that promote the ability to explain one's choices to others and that empower individuals to respond critically to other's reasons. This could be interpreted as a call to restrict deceptive or coercive content. Assessability of content by users should therefore be an important guideline. However, whereas a deliberative view leads engineers to remove mis- and disinformation, thus protecting autonomy with design mechanisms, a liberal view might be more inclined to place this responsibility on the users themselves. Design informed by a liberal view leaves users to decide freely with whom they interact and does not address filter bubbles and echo chambers. Engineers guided by a deliberative view will address filter bubbles and echo chambers and encourage respectful interaction.

Reflecting on these design guidelines and their implications brings to light many questions for further inquiry. While the directive to remove incitement to violence or to remove harmful content and groups seems clear, it necessitates exact definitions of which content incites violence, which content and which users and groups cause mental harm, and which individuals and groups intend to inflict harm. The questions regarding definitions can be difficult to answer; they depend on factors such as context, and determining the exact intentions of users and groups may at times be impossible. Should the act of unintentionally sharing falsehoods be classified as harmful? Where exactly should the lines between free speech, autonomy and equality be drawn? Empirical evidence about the removal of mis- and disinformation and the encouragement of users to engage in debate makes it clear that there is a difference between value embodiment and value realisation through design. Though well

intended, design features may not deliver the values they are designed for. Therefore, engineers should also take into account empirical research about the consequences of specific design choices. Finally, Habermas's guidelines regarding the appropriate ethos of corporations and their engineers and their treatment of users raise questions about how to translate theory to practice and who is responsible for what when designing for a (more) democratic social media.

Finally, we offer some further remarks for future research. First, the design guidelines and solutions presented in this article do not aim to provide a comprehensive overview of what would be fitting to design for a democratic social media. We encourage others to draw out more guidelines and fitting design features. Second, as this paper limits itself to two Western theories of democracy, this review would benefit from an extension. Modern theories of democracy, such as radical democracy theory or modern interpretations of the work of Habermas (such as the work by Simon Chambers) and of the work of Mill, as well as non-Western theories of democracy, could offer contrasting perspectives and prove particularly insightful. This point is emphasised by many of questions raised when translating deliberative and classical liberal democracy theories into the social media context. Finally, in this article, we explicate how varied the appearance and functionality of democratic social media can be, while leaving unanswered the question of responsibility. Who is responsible for designing social media platforms democratically? Yet, as a preceding step, illustrating how a wide range of theories translate into democratic social media design and contemplating the implications of these design choices, as we have done in this paper, can enable social media corporations and their engineers, as well as policymakers, to think more carefully about the online future they are creating.

ACKNOWLEDGMENT

This publication is part of the research program Ethics of Socially Disruptive Technologies (ESDiT), which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

REFERENCES

Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. (2021). Technologically scaffolded atypical cognition: The case of YouTube's recommender system. *Synthese*, *199*(1), 835–858. doi:10.1007/s11229-020-02724-x

Amoretti, M. C., & Lalumera, E. (2019). Harm should not be a necessary criterion for mental disorder: Some reflections on the DSM-5 definition of mental disorder. *Theoretical Medicine and Bioethics*, *40*(4), 321–337. doi:10.1007/s11017-019-09499-4 PMID:31535312

Arneson, R. J. (1980). Mill versus paternalism. *Ethics*, *90*(4), 470–489. doi:10.1086/292179

Bächtiger, A., Dryzek, J. S., Mansbridge, J., & Warren, M. E. (2018). *The Oxford handbook of deliberative democracy*. Oxford University Press. doi:10.1093/oxfordhb/9780198747369.001.0001

Bail, C. (2022). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press., doi:10.1515/9780691246499

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics* (7th ed.). Oxford University Press.

Bell, M. C. (2021). John Stuart Mill's harm principle and free speech: Expanding the notion of harm. *Utilitas*, *33*(2), 162–179. doi:10.1017/S0953820820000229

Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics and Information Technology*, *17*(4), 249–265. doi:10.1007/s10676-015-9380-y

Brough, M., Literat, I., & Ikin, A. (2020). "Good social media?": Underrepresented youth perspectives on the ethical and equitable design of social media platforms. *Social Media + Society*, *6*(2). Advance online publication. doi:10.1177/2056305120928488

Caplan, B. (2001). Rational ignorance versus rational irrationality. *Kyklos*, *54*(1), 3–26. doi:10.1111/1467-6435.00138

Çarçani, K., & Mörtberg, C. (2018). Enhancing engagement and participation of seniors in society with the use of social media: The case of a reflective participatory design method story. *Interaction Design and Architecture(s), 36(SI)*, 58-74

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. doi:10.1016/j.avb.2021.101608

Cohen, J. (2005). Deliberation and democratic legitimacy. In D. Matravers & J. E. Pike (Eds.), *Debates in contemporary political philosophy: An anthology* (pp. 352–370). Routledge.

Cohen-Almagor, R. (2012). Between autonomy and state regulation: J.S. Mill's elastic paternalism. *Philosophy (London, England)*, *87*(4), 557–582. doi:10.1017/S0031819112000411

Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology & Human Values*, *41*(1), 77–92. doi:10.1177/0162243915589635

Dahlberg, L. (2001). The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information Communication and Society*, *4*(4), 615–633. doi:10.1080/13691180110097030

Dahlberg, L. (2011). Re-constructing digital democracy: An outline of four 'positions.'. *New Media & Society*, *13*(6), 855–872. doi:10.1177/1461444810389569

Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, *42*(1), 15–33. doi:10.2478/nor-2021-0002

DiFranco, R. (2020). I wrote this paper for the Lulz: The ethics of internet trolling. *Ethical Theory and Moral Practice*, *23*(5), 931–945. doi:10.1007/s10677-020-10115-x

Doomen, J. (2014). *Freedom and equality as necessary constituents of a liberal democratic state.* [Doctoral dissertation, Leiden University]. https://philarchive.org/rec/DOOFAE

Elder, A. (2020). The interpersonal is political: Unfriending to promote civic discourse on social media. *Ethics and Information Technology*, *22*(1), 15–24. doi:10.1007/s10676-019-09511-4

Farkas, J., & Schou, J. (2019). *Post-truth, fake news and Democracy: Mapping the politics of falsehood*. Routledge. doi:10.4324/9780429317347

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2). Advance online publication. doi:10.37016/mr-2020-009

Fuchs, C. (2015). *Culture and economy in the age of social media*. Routledge. doi:10.4324/9781315733517

Fukuyama, F. (2006). *The end of history and the last man*. Simon and Schuster. Gaozhao, D. (2021). Flagging fake news on social media: An experimental study of media consumers' identification of fake news. *Government Information Quarterly*, *38*(3), 101591. doi:10.1016/j.giq.2021.101591

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *European Economic Review*, *136*, 103772. doi:10.1016/j.euroecorev.2021.103772

Guiora, A., & Park, E. A. (2017). Hate speech on social media. *Philosophia*, *45*(3), 957–971. doi:10.1007/s11406-017-9858-4

Gutmann, A., & Thompson, D. F. (2004). *Why deliberative democracy?* Princeton University Press. doi:10.1515/9781400826339

Habermas, J. (1984). The theory of communicative action: Vol. 1. Reason and the rationalization of society. Beacon Press.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. The MIT Press. doi:10.7551/mitpress/1564.001.0001

Habermas, J. (2022). *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Suhrkamp Verlag.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. doi:10.1177/1745691617702496 PMID:28792862

Hillman, T., Lundin, M., Rensfeldt, A. B., Lantz-Andersson, A., & Peterson, L. (2021). Moderating professional learning on social media—A balance between monitoring, facilitation and expert membership. *Computers & Education*, *168*, 104191. doi:10.1016/j.compedu.2021.104191

Hösle, V. (2004). *Morals and politics*. University of Notre Dame Press.

Huneman, P., & Vorms, M. (2018). Is a unified account of conspiracy theories possible? *Argumenta Oeconomica Cracoviensia*, *3*, 49–72. doi:10.23811/54.arg2017.hun.vor

Kraut, R., Galegher, J., Fish, R., & Chalfonte, B. (2009). Task requirements and media choice in collaborative writing. *Human-Computer Interaction*, *7*(4), 375–407. doi:10.1207/s15327051hci0704_2

Kriplean, T., Toomim, M., Morgan, J., Borning, A., & Ko, A. J. (2012). Is this what you meant?: Promoting listening on the web with reflect. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 1559–1568). IEEE. doi:10.1145/2207676.2208621

Kruger, J., Epley, N., Parker, J., & Ng, Z.-W. (2005). Egocentrism over e-mail: Can we communicate as well as we think? *Journal of Personality and Social Psychology*, *89*(6), 925–936. doi:10.1037/0022-3514.89.6.925 PMID:16393025

Lakier, G. (2021, January 27). *The Great Free-Speech Reversal*. The Atlantic. https://www.theatlantic.com/ideas/archive/2021/01/first-amendment-regulation/617827/

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, *4*(11), 11. doi:10.1038/s41562-020-0889-7 PMID:32541771

MacDermott, Á., Motylinski, M., Iqbal, F., Stamp, K., Hussain, M., & Marrington, A. (2022). Using deep learning to detect social media 'trolls.' *Forensic science international. Digital Investigation*, *43*, 301446. doi:10.1016/j.fsidi.2022.301446

Mansbridge, J. J., Bohman, J., Chambers, S., Estlund, D., Føllesdal, A., Fung, A., Lafont, C., Manin, B., & Luismarti, J. (2011). The place of self-interest and the role of power in deliberative democracy. *Raisons Politiques*, *42*(2), 47–82. doi:10.3917/rai.042.0047

Marin, L., & Roeser, S. (2020). Emotions and digital well-being: The rationalistic bias of social media design in online deliberations. In C. Burr & L. Floridi (Eds.), *Ethics of digital well-being* (Vol. 140, pp. 139–150). Springer International Publishing. doi:10.1007/978-3-030-50585-1_7

Marrone, P. (2022). Epistemic democracy and technopolitics: Four models of deliberation. [IJT]. *International Journal of Technoethics*, *13*(1), 1–14. doi:10.4018/IJT.291551

McKay, S., & Tenove, C. (2020). Disinformation as a threat to deliberative democracy. *Political Research Quarterly*, *74*(1), 106591292093814. doi:10.1177/1065912920938143

Mill, J. S. (1991). *On liberty*. Routledge. (Original work published 1859)

Mouffe, C. (1999). Deliberative democracy or agonistic pluralism? *Social Research*, *66*(3), 745–758.

Mouffe, C. (2005). *The Return of the political*. Verso.

Niemiec, E. (2020). COVID-19 and misinformation. *EMBO Reports*, *21*(11), e51420. doi:10.15252/embr.202051420 PMID:33103289

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you.* Penguin UK.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 7855. Advance online publication. doi:10.1038/s41586-021-03344-2 PMID:33731933

Rauch, J. (2021). *The constitution of knowledge: A defense of truth*. Brookings Institution Press.

Reviglio, U. (2019). Serendipity as an emerging design principle of the infosphere: Challenges and opportunities. *Ethics and Information Technology*, *21*(2), 151–166. doi:10.1007/s10676-018-9496-y

Rienstra, B., & Hook, D. (2006). Weakening Habermas: The undoing of communicative rationality. *Politikon: South African Journal of Political Studies*, *33*(3), 313–339. doi:10.1080/02589340601122950

Rosenberg, M. B., & Chopra, D. (2015). *Nonviolent communication: A language of life: Life-changing tools for healthy relationships*. PuddleDancer Press.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301. doi:10.1016/0022-1031(77)90049-X

Simon, G. (n.d.). *Nudge theory examples in online discussions*. OpenWeb. Retrieved August 16, 2023, from https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. OUP Oxford. doi:10.1093/acprof:oso/9780199580828.001.0001

Somin, I. (2010). Deliberative democracy and political ignorance. *Critical Review: A Journal of Politics and Society, 22*(2-3), 253-279.

Tandoc, E. C. Jr, Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, *21*(3), 381–398. doi:10.1177/1464884919868325

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *The American Economic Review*, *93*(2), 175–179. doi:10.1257/000282803321947001

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thompson, S. (2019). Hate speech and self-restraint. *Ethical Theory and Moral Practice*, *22*(3), 657–671. doi:10.1007/s10677-019-10004-y

Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology*, *22*(1), 69–80. doi:10.1007/s10676-019-09516-z

Valkenburg, G. (2013). Technoethics and public reason. [IJT]. *International Journal of Technoethics*, *4*(2), 72–84. doi:10.4018/jte.2013070106

van de Poel, I. (2021). Design for value change. *Ethics and Information Technology*, *23*(1), 27–31. doi:10.1007/s10676-018-9461-9

Yom-Tov, E., Dumais, S., & Guo, Q. (2014). Promoting civil discourse through search engine diversity. *Social Science Computer Review*, *32*(2), 145–154. doi:10.1177/0894439313506838

Zuiderveen Borgesius, F., Trilling, D., Moeller, J., Bodó, B., de Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review. Journal on Internet Regulation*, *5*(1). Advance online publication.

ENDNOTES

1 https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking.
2 https://communitynotes.twitter.com/guide/en/about/introduction.
3 https://www.inquirer.com/news/twitter-bans-trump-free-speech-first-amendment-20210109.html.
4 See for example the work of McKay and Tenove (2020) on the issue of disinformation; Thompson (2019) and Ullmann and Tomalin (2020) on the issue of hate speech; Marin and Roeser (2020), Reviglio (2019) and Elder (2020) on how to improve civic discourse through design; and Brough, Literat and Ikin (2020) and Çarçani and Mörtberg (2018) on redesigning social media for inclusiveness and equity.
5 Libertarian theories go even further in their defense of free speech. It might be worthwhile in future research to contrast libertarian and liberal theories of democracy and their different conceptualisations of freedom of speech.
6 See, e.g., Arneson (1980), for a discussion of exceptions of Mill' paternalism - such as in the case of children or 'barbarians'. See also Cohen-Almagor (2012).
7 https://www.eu-digital-services-act.com/
8 Once again, a libertarian viewpoint – as opposed to a liberal one – might go even further and remove only incitements to violence, but not other harmful or deceptive content.
9 As an anonymous reviewer pointed out, respect for equality as such is a principle that most democracy theories would subscribe to; hence this principle would also find support from a classical liberal viewpoint.
10 https://www.sciencedaily.com/releases/2019/09/190916092101.htm
11 https://apnews.com/article/election-2020-joe-biden-donald-trump-capitol-siege-media-e79eb5164613d6718e9f4502eb471f27
12 https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api
13 We must note, however, that the idea of the existence of filter bubbles is challenged and even rejected by some authors (see, e.g., Dahlgren, 2021).
14 https://blog.twitter.com/en_us/topics/product/2020/new-conversation-settings-coming-to-a-tweet-near-you
15 https://ilp.mit.edu/node/43998

Roxanne van der Puil is a researcher at Eindhoven University of Technology with the Philosophy & Ethics research group. Her PhD dissertation focuses on social media and democracy, with a specific focus on post-truth and design.

Andreas Spahn is an associate professor with the Philosophy& Ethics research group at the department of Industrial Engineering & Innovation Sciences. His research focuses on ethics of technology, ethics of behaviour change technologies, ethics of energy systems, and environmental ethics.

Lambèr Royakkers is professor Ethics of the Digital Society with the Philosophy & Ethics research group at the Eindhoven University of Technology. His main areas of research are applied ethics, ethical theory, and moral responsibility. More specifically, he has recently published on social robots, cyberwar, and the ethics of digitalisation.